Introduction to statistics:

Abdisa Gurmessa (PhD)
(Ass. Prof in Statistics, Specialization in Biostatistics)

Emerald International College,

For MSc in Data Science, Ethiopia

Outlines

1. Introduction to statistics
    - Definition of statistics
    - The importance of statistics
    - Scale of measurement
    - Types of data
2. Methods of data Collection and presentation
    - Methods of data Collection
    - Method of Data Organization and Presentation
        - Tabular methods
        - Graphical methods
3. Measures of Central Tendency and Dispersion
    - Methods of data summarization
        - Measures of Central Tendency
        - Measure of dispersion

Outlines

4. Probability theory and Probability Distribution
    - Definition and rules of probability
    - Permutation
    - Combination
    - Random Variables and Probability Distribution
    - Discrete Probability Distribution
        - Binomial distribution
        - Poisson distribution
    - Continuous Probability Distribution
        - Normal distribution
5. Sampling Method and Sample Size Calculation
    - Sampling Methods
        - Probability Sampling Method
        - Nonprobability Sampling Method
    - Sample size Calculation
6. Basic Concepts of Inference
    - Point Estimation
    - Interval Estimation
    - Test of hypothesis

Chapter One

1. Introduction to Statistics

What Is Statistics?

Example: Sickness absence

- In occupational medicine, one is interested in studying factors that influence absence due to sickness
- The following data were obtained from 585 employees with a similar job:

Gender	Sickness absence		Total
	No	yes	
Female	245	184	429
Male	98	58	156
Total	343	242	585

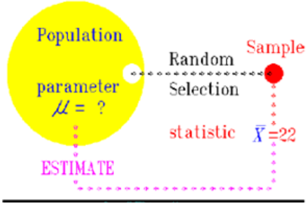
Research question

Is there a relation between absence and gender ?

- $184/429 = 42.9\%$ of the females, and $58/156 = 37.2\%$ have been absent.
- This suggests that females are more absent than males
- However, even if absence due to sickness is equally frequent amongst males and females, the above results could have occurred by pure chance.
- Therefore it is important to calculate how likely it would be to observe such differences, by pure chance
- If this would not be unlikely, then the data provide no evidence for such a relation.

Conclusion

- The aim of statistics is twofold:
  - Descriptive statistics: Summarizing and describing observed data such that the relevant aspects are made explicit.
  - Inferential statistics: Studying to what extent observed trends/effects can be generalized to a general (infinite) population.



Some Definition

- Statistics: The development and application of theory and methods to the Collection, organization, presentation, analysis and interpretation of data and
drawing of inferences about a body of data when only a small part of the data is observed.
- Data science: is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data.
- This broad field includes the development of statistical theory and methods, and their applications in different descpline
- Data scienctistis play essential roles in designing studies, analyzing data and creating methods to attack research problems as diverse as:

Information (data) is collected to investigate a real life problem

The methods and tools of statistics are used to analyze the data for decision making.

The discipline of statistics consists of different tools which are applicable in many fields including health, agriculture, business, economics, education, psychology etc.

Classification of Statistics

Depending on how data can be used, statistics can be divided into two main areas or branches.

Descriptive statistics: Statistical method that deals with the process organizing or summarizing a given set of data into a meaningful form.

In descriptive statistics there is no generalization or conclusion about the population.

It consists of collection, organization and presentation of data in the form of tables, graphs and numerical summary measures (measure of central tendecy and measure of variability)

Example: prevalence, incidence, mean, median, etc.

Inferential Statistics

Is the process of drawing conclusion (inference) about a population based on the information obtained from the sample.

Consists of performing hypothesis testing, determining relationships among variables and making predictions.

The major contribution of statistics is that it enables us to use data from the sample to make estimates and test claims about the characteristics of a population.

- Population: The entire ”universe” of possibilities.
- It is the group of people from which a statistical sample is taken in statistics.

- Sample: A part of the population that we can observe.

Population versus sample

- Observed data can always be considered as taken from some population.

- The strength of the evidence in the data, and the validity of the conclusions based on the data depends entirely on:

  - The definition of the population

  - The way the sample is drawn from the population.

- Parameter: A descriptive measure computed from the data of a population.
- Statistic: A descriptive measure computed from the data of a sample.

Types of Variables and Measurement scales

Variable: A variable is a characteristics or attribute that can take on different possible value or outcome.

Example Blood pressure, enzyme level, height, weight, sex, salary, etc

A variable whose values are determined by measurement, counting or by chance.

Variable and Types of Data

The measurement or observation value of a variable is called a data and collection of data is known as data set

A variable can be qualitative or quantitative

Qualitative or Categorical variable: A variable or characteristic which can not be measured in quantitative(numerical) form but can only be sorted (or grouped) by name or categories.

Quantitative (Numerical) Variable: A variable that can be measured (or counted) and expressed numerically.

E.g. Observations regarding height, income, weight, age, etc...

A quantitative variable is divided into two

Discrete Variable : A variable that can assume only certain counted values and for which there is gaps between any two possible values.

The values aren't just labels, but are actual measurable quantities.

Example Number of students in class, Number of tree species in Ethiopia , Number of children/ family, number of bacteria colonies on a plate.

Continuous Variable: : It can have an infinite number of possible values in any given interval. There is no gap between any two possible values.

Example All variable whose values obtained through measurement like height, weight, length

Scales of measurement

Scales of measurement refers to ways in which variables or numbers are defined and categorized.

The various measurement scales results from the facts that measurement may be carried out under different sets of rules.

There are four types of scales of measurement.

A) Nominal scale: The type of data, in which the values fall into unordered categories or classes.

It consists of ’naming’ observations or classifying them into various mutually exclusive categories.

Example Gender, blood group, Martial status, Political affiliation, Eye color.

B) Ordinal Scale: - when the observation category represents an ordered series of relationship or rank order.

The variables deal with their relative difference rather than with quantitative differences.
Ordinal data are data which can have meaningful inequalities.

Example Letter grading system, level of education, Military status, Individuals may be classified according to socio-economic.

Note Nominal and ordinal scale of measurement are classified under categorical variable

C) Interval Scale: A scale of measurement represents quantity and has equal units but for which zero represents simply an additional point of measurements.

The distance between any two measurements is known but not meaningful quotients.

Example Temperature (C/F), income, IQ measurement

D) Ratio Scale: Characterized by the fact that equality of ratios as well as equality of intervals may be determined.

Measurement begins at a true zero point and the scale has equal space.

Ratio between any two measurement is meaningful

Example most physical quantities like length, mass, weight, Social variable like age, length of residence in a given place.

Note: Interval and ratio scale of measurement are classified under quantitative variable

Chapter Two

2. Methods of Data Collection and Presentation

2. Methods of Data Collection and Presentation

Data are numbers which are values of measurements or counting of a variable

- There are various ways data can be collected:
    - Primary source : First hand information is gathered
      Data measured or collected by the investigator or the user directly from the source
    - Secondary source : Documented
      which were collected by some other agency before and it may be published or unpublished.
        - They are less expensive in time and cost than Primary data.
        - Usually they are published or unpublished materials, records, reports, e t c.
    - Data which are primary for one may be secondary for the other.
    - The strength of evidence depends on the method of data collection.

Method of Data Collection

Having decided on how to design the research study, the next methodological design is how to collect information.

Depending on the type of variable and the objective of the study different data collection methods can be employed.

The choice of methods of data collection is based on:

- The accuracy of information they will yield
- Practical considerations, such as, the need for personnel, time, equipment and other facilities, in relation to what is available.

The most commonly used methods of collecting information (quantitative data) are the use of documentary sources, interviews and self-administered questionnaires.

Document Review

Advantage

- Relatively inexpensive
- Good source of background information
- Provides information may not be directly observable
- May bring up issues not noted by other means

Disadvantage

- Information may be inapplicable, disorganized, unavailable or out of date.
- Information may be incomplete or inaccurate
- Can be time consuming to collect, review, and analyze many documents

Interviews

Advantage

- Useful for gaining insight and context into a topic
- Allows respondents to describe what is important to them
- Useful for gathering quotes and stories

Disadvantage

- Susceptible to interview bias
- Time consuming and expensive compared to other data collection methods
- May seem valueless to the respondent

Questionnaires

Advantage

- Administration is comparatively inexpensive
- Easy even when gathering data from large numbers of people spread over wide geographic area
- Reduces chance of evaluator bias because the same questions are asked of all respondents
- Some people feel more comfortable responding to a survey than participating in an interview
- Tabulation of closed-ended responses is an easy and straightforward process

Disadvantage

- Survey respondents may not complete the survey resulting in low response rates
- Items may not have the same meaning to all respondents
- Size and diversity of sample will be limited by people's ability to read
- Unable to contact for additional details
- Good survey questions are hard to write and they take considerable time to develop and refine.

In questionnaire design remember to:

Pretest the questionnaire on 20-50 respondents in actual field situation

Use familiar and appropriate language

Avoid abbreviations, double negatives, etc

Arrange questions in logical sequence

Start with simpler question

Check all filled questionnaire at field level

Methods of collecting qualitative data

Qualitative approaches to data collection usually involve direct interaction with individuals on a one to one basis or in a group setting.

Focus Groups

Advantage

- Quick and relatively easy to set up
- Group dynamics can provide useful information that individual data collection does not provide
- Is useful in gaining insight into a topic that may be more difficult to gather information through other data collection methods

Disadvantage

- Susceptible to facilitator bias
- Discussion can be dominated or side tracked by a few individuals
- Data analysis is time consuming and needs to be well planned in advance
- Does not provide valid information at the individual level
- The information is not representative of other groups

Observational

Advantage

- Collect data where and when an event or activity is occurring
- Does not rely on people's willingness to provide information
- Directly see what people do rather than relying on what they say.

Disadvantage

- Susceptible to observer bias
- People usually perform better when they know they are being observed
- Does not increase understanding of why people behave the way they do.

When should you use observation for evaluation?

When you are trying to understand an on going process or situation.

When you are gathering data on individual behaviors or interactions between people.

When you need to know about a physical setting.

When data collection from individuals is not a realistic option

How do you plan for observations?

Determine the focus topic

Design a system for data collection.

Recording sheets and checklists

Select the sites.

Select the observers.

Train the observers.

Time your observations appropriately

Method of Data Organization and Presentation

The data collected in statistical studies are often so large that must be reduced to a manageable proportions before any study of them can be begin or can be made.

The data collected in a survey is called raw data.

For the primary objective of data organization and presentation arrays, tables and diagrams are commonly used.

Tabular Presentation (Frequency Distribution)

Frequency: - is the number of times a certain value or set of values occurs in a specific group.

A frequency distribution: is a table that shows data classified into a number of classes with a corresponding number of times falling in each class (frequency)

Classification: - is the process of arranging items/data into classes or categories according to their similarities and/or differences.

depending up on the nature of the data in hand frequency distribution can be defined as qualitative and quantitative frequency distribution.

Quantitative frequency distribution can be further classified as ungrouped and grouped frequency distribution.

Example: let as assume that 20 women were surveyed to find out how many children they had. The raw data obtained from the survey is as follow:
0, 2, 3, 1, 1, 3, 4, 2, 4, 2, 2, 1, 0, 4, 1, 2, 4, 2, 3, 4.

The ungrouped frequency distribution for the data is given below.

Number of children	Frequencyy
0	2
1	4
2	6
3	3
4	5
Total	20

Relative frequency: is a value obtained by dividing the frequencies in each class of the frequency distribution by the total number of observations.

Mathematically, , where “f_i” is the frequency of each class and “n” is the total number of observations.

$$r.f = \frac{f_i}{n}$$

To find the percentage of each value we multiply the relative frequency by 100%. And the percentage can be formulated as:-

$$\text{percentage\%} == \frac{f_i}{n} \times 100\%$$

For the above data (women and their number of children), the relative frequency and the percentage become:

Number of children	Frequencyy	Rel.freq	Percent(%)
0	2	0.1	10
1	4	0.2	20
2	6	0.3	30
3	3	0.15	15
4	5	0.25	25
Total	20	1	100

Cumulative frequency distribution: Shows the number of units lie below or above the specified unit or class of interval with inclusive.

When the interest of the investigator is on the number of cases below the specified value (the upper limit of the interval), it is known as "less than" cumulative frequency distribution.

When the interest lies in finding the number of cases above a specified value, then this value (lower limit of the specified interval), then, it is known as "more than" cumulative frequency distribution.

E.g : For the above data (women and their number of children) data, we can see that:-

Number of children	Frequencyy	"less" than cumulative freq	"More than" cumulative frequency
0	2	2	20
1	4	6	18
2	6	12	14
3	3	15	8
4	5	20	5
Total	20		

Grouped frequency distribution with class intervals

To construct interval frequency distributions, the following points are mandatory:

1. The number classes should be clearly defined
   A guide on the determination of the number of classes (k) can be the Sturge's Formula, given by: $K = 1 + 3.322 \times log(n)$, where n is the number of observations.
   Commonly, with the understanding of the number of classes should not be too large or too small. It is recommended to be among 5 to 15 in accordance of the number of observations considered.
2. All intervals should be with the same width. After fixing the number of intervals (classes), the width of the interval is found using

$$\text{The width of the interval} = \frac{\text{Range}}{\text{number of classes}} \Rightarrow W = R/K;$$

where Range =maximum value-minimum value

Example: The number of hours of 25 emergency patients spent in emergency room at a given hospital is given below. Construct a suitable frequency distribution for these data using 5 classes.

62	50	35	36	31
41	31	65	30	41
37	62	27	47	65
27	53	40	29	63
58	65	38	41	26

Step 1: Max = 65, Min = 26, so the range; R = 65-26 = 39.
Step 2 : It is already determined to construct a frequency distribution having 5 classes
Step 3: Class width: $W = 39/5 = 7.8 \approx 8$
Step 4: Starting point = 26 = lower limit of the first class.
And hence the lower class limits become: 26 34 42 50 58
Step 5 : Upper limit of the first class = 34-1 = 33.
And hence the upper class limits become: 33 41 49 57 65

The class limits of the above ungrouped data and the frequency distribution results of the steps were listed in the following table.

Class limits	frequency
26-33	7
34-41	8
42-49	1
50-57	2
58-65	7

Exercise: The following data indicates the amount of a given enzyme concentration of 30 women during pregnancy:

i. Construct a grouped frequency distribution

ii. Find the relative frequency, percentage, the less than and more than cumulative frequency for each class interval

23	24	18	14	20	24	24	27	34	21
15	16	18	23	22	21	30	29	30	20
16	28	12	27	23	26	19	32	34	37

Two-way table (Cross tabulation):

This table shows two characteristics and is formed when either of the two variables (the caption or the stub) is divided into two or more parts.

For instance , the marital status and cervical cancer status can be presented in the following two way table.

Marital status	Cervical Cancer status		Total
	Positive	Negative	
Single	49	47	96
Married	216	108	324
Widowed	87	86	xxx
Div/sep	15	45	xxx
Total	367	XXX	XXX

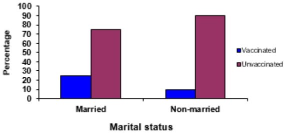
Graphical (Diagrammatic) Presentation of Data

I. Bar Graph

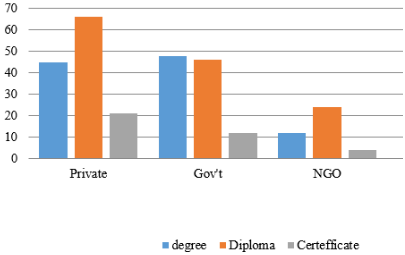
The bar graph is very commonly used and is better for representation of qualitative data. Bars are vertical lines, where the lengths of the bars are proportional to their corresponding numerical values and the bars should be equally space.

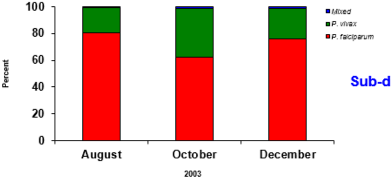
Example: if following data indicates the number clinical Nurses in given woreda, it can be presented using different diagrams.

	Degree	Diploma	Certefficate
Private	45	66	21
Gov't	48	46	12
NGO	12	24	4



Multiple bar graph





Sub-divided bar graph

II. Histogram

Histograms are frequency distributions with continuous class intervals that have been turned into graphs.

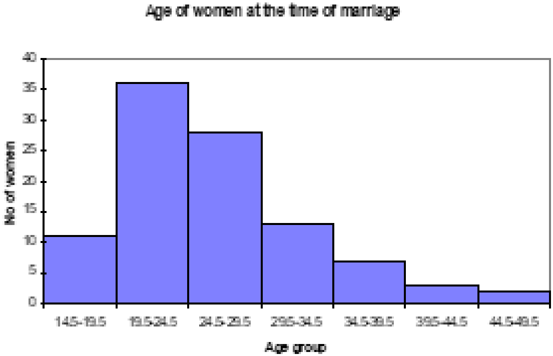
To construct a histogram, we draw the interval boundaries on a horizontal line and the frequencies on a vertical line.

Non-overlapping intervals that cover all of the data values must be used.

Bars are drawn over the intervals in such a way that the areas of the bars are all proportional in the same way to their interval frequencies.

Example: Distribution of the age of women at the time of marriage

Age group	15-19	20-24	25-29	30-34	35-39	40-44	45-49
Number	11	36	28	13	7	3	2


Age of women at the time of marriage
No of women
0
5
10
15
20
25
30
35
40
14.5-19.5
19.5-24.5
24.5-29.5
29.5-34.5
34.5-39.5
39.5-44.5
44.5-49.5
Age group

III. Pie diagram (Pie chart)

Pie chart enables us to show the partitioning of a total in to its component parts.

The diagram is in the form of circle and component as slices of the circle.

The size of the slice represents the proportion of the component out of the total.

The angle of a component (x) is calculated as:

$$\text{Degree of X} = \left(\frac{\text{value of component X}}{\text{total value of the components}} \right) \times 360^{0}$$

Example: The following data indicates the marital status of 40 women who came for the service of contraceptives to St. Paul HMMC. Present the data using Pie-diagram.

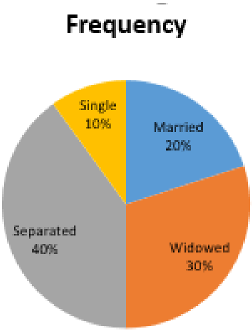
Marital status	married	Widowed	separated	single
Frequency	8	12	16	4

Degree of the slice for married is calculated as:

$$\text{Degree of Married women} = \left(\frac{\text{number of mrried women}}{\text{total women}}\right) \times 360^0$$

$$\text{Degree of Married women} = \left(\frac{8}{40}\right) \times 360^0 = 72^0$$

Like with the slice degree of the pie chart of the women for widowed, separated and single women becomes is 108, 144 and 36, respectively.



IV. The line diagram

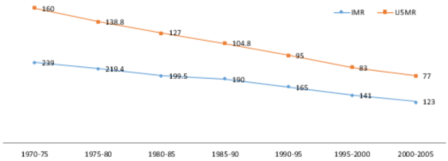
The line graph is especially useful for the study of some variables according to the passage of time.

The time, in weeks, months or years is marked along the horizontal axis; and the value of the quantity that is being studied is marked on the vertical axis.

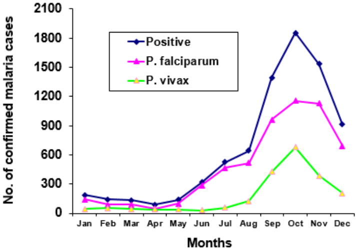
The distance of each plotted point above the base-line indicates its numerical value. The line graph is suitable for depicting a consecutive trend of a series over a long period.

Example: Infant and under five mortality rate in Ethiopia, 1970-2005 (Tefera Darge 2011; EDHS, 2000, 2005)

	1970-75	1975-80	1980-85	1985-90	1990-95	1995-2000	2000-05
IMR	239	219.4	199.5	190	165	141	123
U5MR	160	138.8	127	104.8	95	83	77



Infant and under five mortality rate in Ethiopia, 1970-2005 (Tefera Darge 2011; EDHS, 2000, 2005)



No of microscopically confirmed malaria cases by species and month at Zeway malaria control unit, 2003

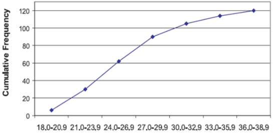
V. Frequency Polygon

If we join the midpoints of the tops of the adjacent rectangles of the histogram with line segments a frequency polygon is obtained.

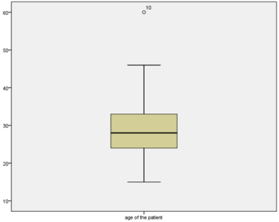
When the polygon is continued to the X-axis just outside the range of the lengths the total area under the polygon will be equal to the total area under the histogram.

Example :Table : Body Mass Index (BMI) Data (n = 120)

Class Interval For BMI Levels	Frequency (f)	Cumulative frequency (cf)
18.0-20.9	6	6
21.0-23.9	24	30
24.0-26.9	32	62
27.0-29.9	28	90
30.0-32.9	15	105
33.0-35.9	9	114
36.0-38.9	6	120

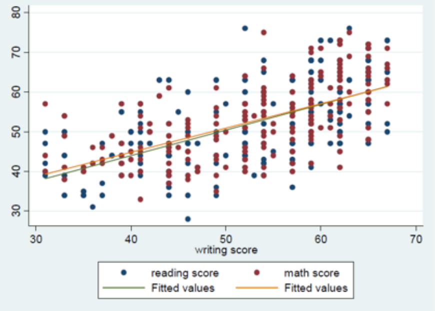


The cumulative frequency polygon (Ogive) presentation for Body Mass Index (BMI) of 120 adults presented in the above table.



An example of box plot graph

An example scatter plot graph



Reading Assignment: read about the following Biostatistical presentation of statistical data and their interpretation:

Box –plot

stem and leaf

scatter plot

Chapter Three

3. Measures of central tendency and dispersion

Measures of central tendency and dispersion

Before attemting the measure of central tendency and dispersion, let's see some of the notations that are used frequency.

$$\text{Notations : } \quad \sum = \text{summation}$$

$$\mu = \text{the mean of the population}$$

$$\sigma = \text{standard deviation of the population}$$

Suppose n values of a variable are denoted as $X_1, X_2, \cdots, X_n$ then
$\sum X_i = X_1 + X_2 + X_3 + ... + X_n$

Property of Summation Notation

Measures of central tendency and dispersion

Before attemting the measure of central tendency and dispersion, let's see some of the notations that are used frequency.

$$\text{Notations :} \quad \sum = \text{summation}$$

$$\mu = \text{the mean of the population}$$

$$\sigma = \text{standard deviation of the population}$$

Suppose n values of a variable are denoted as $X_1, X_2, \cdots, X_n$ then $\sum X_i = X_1 + X_2 + X_3 + ... + X_n$

Property of Summation Notation

1. $\sum_{i=1}^{n} K = nK$, where K is any constant.
2. $\sum_{i=1}^{n} KX_i = K\sum_{i=1}^{n} X_i$, where K is any constant.
3. $\sum_{i=1}^{n} (\alpha + \beta X_i) = n\alpha + \beta\sum_{i=1}^{n} X_i$, where α and β is any constants.
4. $\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$

Measures of central tendency

- Measures of central tendency or measures of location provide us with a summary that describes some central or middle point of the data.
- Measure of central tendency is concerned only with quantitative variables and is undefined for qualitative variables as these are immeasurable on a scale.

Measures of central tendency

- Measures of central tendency or measures of location provide us with a summary that describes some central or middle point of the data.
- Measure of central tendency is concerned only with quantitative variables and is undefined for qualitative variables as these are immeasurable on a scale.

Objective

- To facilitate comparison between two or more different population
- Reduce the bulk of data (i.e condensation of data set)
- To make further statistical analysis

There are several different measures of central tendency.

Measures of central tendency

- Measures of central tendency or measures of location provide us with a summary that describes some central or middle point of the data.
- Measure of central tendency is concerned only with quantitative variables and is undefined for qualitative variables as these are immeasurable on a scale.

Objective

- To facilitate comparison between two or more different population
- Reduce the bulk of data (i.e condensation of data set)
- To make further statistical analysis

There are several different measures of central tendency.

1. The Mean (Arithmetic, Geometric and Harmonic)
2. The mode
3. The Median
4. The Quintiles (Quartiles, Deciles and Percentiles)

Arithmetic Mean

The arithmetic mean, usually abbreviated to A.M 'mean' is the sum of all observation divided by the number of observations.

The mean is usually denoted by $\overline{X}$ and can be calculated for grouped and ungrouped data

The ungrouped data type

Arithmetic Mean

The arithmetic mean, usually abbreviated to A.M ‘mean’ is the sum of all observation divided by the number of observations.

The mean is usually denoted by $\overline{X}$ and can be calculated for grouped and ungrouped data

The ungrouped data type

If X is a variable which takes a values $x_1, x_2, x_3, \ldots, x_n$, in a samples of size n then the arithmetic mean is defined as the total number of observation divided by its sample size. i.e

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n} \tag{1}$$

Arithmetic Mean

The arithmetic mean, usually abbreviated to A.M 'mean' is the sum of all observation divided by the number of observations.

The mean is usually denoted by $\overline{X}$ and can be calculated for grouped and ungrouped data

The ungrouped data type

If X is a variable which takes a values $x_1, x_2, x_3, \ldots, x_n$, in a samples of size n then the arithmetic mean is defined as the total number of observation divided by its sample size. i.e

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n} \tag{1}$$

Example 1: Suppose a student scores on seven examinations were 10, 15, 30, 7, 42 , 79 and 83 ,find the arithmetic mean of a student scores
Solution:

Arithmetic Mean

The arithmetic mean, usually abbreviated to A.M 'mean' is the sum of all observation divided by the number of observations.

The mean is usually denoted by $\overline{X}$ and can be calculated for grouped and ungrouped data

The ungrouped data type

If X is a variable which takes a values $x_1, x_2, x_3, \ldots, x_n$, in a samples of size n then the arithmetic mean is defined as the total number of observation divided by its sample size. i.e

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n} \tag{1}$$

Example 1: Suppose a student scores on seven examinations were 10, 15, 30, 7, 42 , 79 and 83 ,find the arithmetic mean of a student scores
Solution:

$$\overline{x} = \frac{\sum_{i=1}^{7} x_i}{n} = \frac{10 + 15 + 30 + 7 + 42 + 79 + 83}{7} = \frac{266}{7} = 38$$

For ungrouped frequency distribution

let x_1 occur f_1 times ... x_n times f_n in the data set. Then arithmetic mean is given by

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \tag{2}$$

For ungrouped frequency distribution

let x_1 occur f_1 times ... x_n times f_n in the data set. Then arithmetic mean is given by

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \tag{2}$$

where x_i is the i^{th} class observation
k is the number of class
f_i is the i^{th} class frequency

For ungrouped frequency distribution

let x_1 occur f_1 times ... x_n times f_n in the data set. Then arithmetic mean is given by

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \tag{2}$$

where x_i is the i^{th} class observation
k is the number of class
f_i is the i^{th} class frequency

Example 2: Calculate the mean for the following age distribution.

X_i	2	3	7	8	Total
f_i	2	1	3	1	7
$f_i X_i$	4	3	21	8	36

Solution:

For ungrouped frequency distribution

let x_1 occur f_1 times ... x_n times f_n in the data set. Then arithmetic mean is given by

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \tag{2}$$

where x_i is the i^{th} class observation
k is the number of class
f_i is the i^{th} class frequency

Example 2: Calculate the mean for the following age distribution.

X_i	2	3	7	8	Total
f_i	2	1	3	1	7
$f_i X_i$	4	3	21	8	36

Solution:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{36}{7} = 5.14$$

Grouped data set

The procedure for finding arithmetic mean for grouped data set is similar to the ungrouped data set with frequency, except that the mid points of the class are used as x values.

Grouped data set

The procedure for finding arithmetic mean for grouped data set is similar to the ungrouped data set with frequency, except that the mid points of the class are used as x values.

Let x_i be the mid point of the i^{th} class for i=1, 2, 3, ,..., K having the corresponding frequency f_i then

Grouped data set

The procedure for finding arithmetic mean for grouped data set is similar to the ungrouped data set with frequency, except that the mid points of the class are used as x values.

Let x_i be the mid point of the i^{th} class for i=1, 2, 3, ,..., K having the corresponding frequency f_i then

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} \tag{3}$$

where
x_i is the mid point of i^{th} class
k is the number of class
f_i is the i^{th} class frequency

Example 3: calculate the mean for the following age distribution

Example 3: calculate the mean for the following age distribution

Class	6-10	11-15	16-20	21-25	26-30	31-35
frequency	35	23	15	12	9	6

Solution: First find the class marks

Find the product of frequency and class marks
Finally, find mean using the formula.

Example 3: calculate the mean for the following age distribution

Class	6-10	11-15	16-20	21-25	26-30	31-35
frequency	35	23	15	12	9	6

Solution: First find the class marks

Find the product of frequency and class marks
Finally, find mean using the formula.

Class	6-10	11-15	16-20	21-25	26-30	31-35	Total
f_i	35	23	15	12	9	6	100
X_i	8	13	18	23	28	33	
$X_i f_i$	280	299	270	276	252	198	1575

$$\overline{X} = \frac{\sum_{i=1}^{6} f_i X_i}{\sum_{i=1}^{6} f_i} = \frac{1575}{100} = 15.75$$

properties of Arithmetic mean

- The sum of the deviations of a set of items from their mean is always zero. i.e.

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

- The sum of the squared deviations of a set of items from their mean is the minimum. i.e.

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 < \sum_{i=1}^{n}(X_i - A)^2, A \neq \overline{X}$$

- Easy to calculate and understand (simple).
- Calculation based on all observations
- Not much affected by sampling fluctuations
- Useful for further algebraic calculations
- Useful for comparison purposes
- Greatly affected by the extreme values.
- In case of grouped data if any class interval is open, arithmetic mean can not be calculated.

2. Median

In a distribution, median is the value of the variable which divides it in to two equal halves.

It is the middle most value in the sense that the number of values less than the median is equal to the number of values greater than it.

If the number of observations is odd, the median will be the middle value when all values are arranged in order of magnitude.

When the number of observations is even, there is no single middle value but two middle observations.

In this case the median is the mean of these two middle observations, when all observations have been arranged in the order of their magnitude.

Median is denoted by $\widetilde{X}$

If $X_1, X_2, ..., X_n$ be the observations, then the numbers arranged in ascending order will be $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$, where $X_{[i]}$ is i^{th} smallest value. $X_{[1]} < X_{[2]} <, ..., < X_{[n]}$ then

If $X_1, X_2, ..., X_n$ be the observations, then the numbers arranged in ascending order will be $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$, where $X_{[i]}$ is i^{th} smallest value. $X_{[1]} < X_{[2]} <, ..., < X_{[n]}$ then

$$\widetilde{X} = \begin{cases} X_{[\frac{(n+1)}{2}]}, & \text{if n is odd} \\ \frac{1}{2}\left(X_{[\frac{n}{2}]} + X_{[\frac{n}{2}+1]}\right), & \text{if n is even} \end{cases}$$

Example: Find the median of the following numbers.

If $X_1, X_2, ..., X_n$ be the observations, then the numbers arranged in ascending order will be $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$, where $X_{[i]}$ is i^{th} smallest value. $X_{[1]} < X_{[2]} <, ..., < X_{[n]}$ then

$$\widetilde{X} = \begin{cases} X_{[\frac{(n+1)}{2}]}, & \text{if n is odd} \\ \frac{1}{2}\left(X_{[\frac{n}{2}]} + X_{[\frac{n}{2}+1]}\right), & \text{if n is even} \end{cases}$$

Example: Find the median of the following numbers.

a. 6, 5, 2, 8, 9, 4.

b. 2, 1, 3, 5, 8.

Solution:

If $X_1, X_2, ..., X_n$ be the observations, then the numbers arranged in ascending order will be $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$, where $X_{[i]}$ is i^{th} smallest value. $X_{[1]} < X_{[2]} <, ..., < X_{[n]}$ then

$$\widetilde{X} = \begin{cases} X_{[\frac{(n+1)}{2}]}, & \text{if n is odd} \\ \frac{1}{2}\left(X_{[\frac{n}{2}]} + X_{[\frac{n}{2}+1]}\right), & \text{if n is even} \end{cases}$$

Example: Find the median of the following numbers.

a. 6, 5, 2, 8, 9, 4.

b. 2, 1, 3, 5, 8.

Solution:

a. Order the data set
2,4,5,6,8,9 , and $n = 6$ is even

$$\widetilde{X} = \frac{1}{2}\left(X_{[\frac{6}{2}]} + X_{[\frac{6}{2}+1]}\right) = \frac{1}{2}(5+6) = 5.5$$

b. 1, 2, 3, 5, 8, and $n = 5$ is odd. Hence $\widetilde{X} = X_{\frac{5+1}{2}} = X_3 = 3$

Median for grouped data

In calculating the median from grouped data, we assume that the values within a class-interval are evenly distributed through the interval.

The first step is to locate the class interval in which the median is located, using the following procedure.

Find n/2 and see a class interval with a minimum cumulative frequency which contains n/2. Then the median is obtained from the median class as

Median for grouped data

In calculating the median from grouped data, we assume that the values within a class-interval are evenly distributed through the interval.

The first step is to locate the class interval in which the median is located, using the following procedure.

Find n/2 and see a class interval with a minimum cumulative frequency which contains n/2. Then the median is obtained from the median class as

$$\widetilde{X} = L_{med} + \frac{w}{f_{med}}\left(\frac{n}{2} - c\right) \tag{4}$$

where L_{med} = the lower class boundary of the median class
w= the size of the median class
n = total number of observations
c = the cumulative frequency (less than type) preceeding the median class
f_{med} = frequency of the median class

Note: The median class is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{n}{2}$

Example: Find the median of the following distribution.

Size of farm	No of Farms	Cum.Freq($\leq$)
4.5-14.5	8	8
14.5-24.5	12	20
24.5-34.5	17	37
34.5-44.5	29	66
44.5-54.5	31	97
54.5-64.5	5	102
64.5-74.5	3	105

Solutions:

Example: Find the median of the following distribution.

Size of farm	No of Farms	Cum.Freq($\leq$)
4.5-14.5	8	8
14.5-24.5	12	20
24.5-34.5	17	37
34.5-44.5	29	66
44.5-54.5	31	97
54.5-64.5	5	102
64.5-74.5	3	105

Solutions:Now the median class is at $\frac{n}{2}^{th}$ in the data set.
$\frac{n}{2} = \frac{105}{2} = 52.5$ which occurs in 66 frequency of the less than type cumulative frequency.
$\Rightarrow$ median class = 34.5-44.5, $l_{med} = 34.5$, $f_{med} = 29$, $w = 10$, $c = 37$

Example: Find the median of the following distribution.

Size of farm	No of Farms	Cum.Freq($\leq$)
4.5-14.5	8	8
14.5-24.5	12	20
24.5-34.5	17	37
34.5-44.5	29	66
44.5-54.5	31	97
54.5-64.5	5	102
64.5-74.5	3	105

Solutions:Now the median class is at $\frac{n}{2}^{th}$ in the data set.
$\frac{n}{2} = \frac{105}{2} = 52.5$ which occurs in 66 frequency of the less than type cumulative frequency.
$\Rightarrow$ median class = 34.5-44.5, $l_{med} = 34.5$, $f_{med} = 29$, $w = 10$, $c = 37$

$$\widetilde{X} = L_{med} + \frac{w}{f_{med}}\left(\frac{n}{2} - c\right) = 34.5 + \frac{10}{29}\left(\frac{105}{2} - 37\right) \approx 39.845$$

Properties of the median

There is only one median for a given set of data (uniqueness)

The median is easy to calculate

Median is a positional average and hence it is insensitive to very large or very small values

Median can be calculated even in the case of open end intervals

It is determined mainly by the middle points and less sensitive to the remaining data points (weakness).

It is not a good representative of data if the number of items is small

3. Mode

Mode is a value which occurs most frequently in a set of values

Any observation of a variable at which the distribution reaches a peak is called a mode.

It is not influenced by extreme values.

3. Mode

Mode is a value which occurs most frequently in a set of values

Any observation of a variable at which the distribution reaches a peak is called a mode.

It is not influenced by extreme values.

The set of data may not have mode or may have one mode (unimodel) or two model (bimodel) or more than two mode (multi-model)

It is not a good summary ofor the majority of the data.

The mode of a set of numbers $X_1, X_2, ..., X_n$ is usually denoted by $\widehat{X}$

3. Mode

Mode is a value which occurs most frequently in a set of values

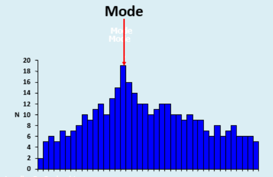
Any observation of a variable at which the distribution reaches a peak is called a mode.

It is not influenced by extreme values.

The set of data may not have mode or may have one mode (unimodel) or two model (bimodel) or more than two mode (multi-model)

It is not a good summary ofor the majority of the data.

The mode of a set of numbers $X_1, X_2, ..., X_n$ is usually denoted by $\widehat{X}$



Ungrouped data

It is a value which occurs most frequently in a set of values.

If all the values are different there is no mode, on the other hand, a set of values may have more than one mode.

$\widehat{X}$ = The most frequently value in the data set
ample: Find the mode of

A. 5, 3, 5, 8, 9

B. 8, 9, 9, 7, 8, 2, 5

C. 4, 12, 3, 6, 7

Solutin:

Ungrouped data

It is a value which occurs most frequently in a set of values.

If all the values are different there is no mode, on the other hand, a set of values may have more than one mode.

$\widehat{X}$ = The most frequently value in the data set
ample: Find the mode of

A. 5, 3, 5, 8, 9

B. 8, 9, 9, 7, 8, 2, 5

C. 4, 12, 3, 6, 7

Solutin:

A Mode, $\widehat{X}$ =5 (unimodel)

B. Mode, $\widehat{X}$ = 8 and 9 (Bimodel)

C. $\widehat{X}$ = No mode for the data.

grouped data

The mode for the grouped data can be obtained from the modal class

Model class is a class having the largest frequency.

The mode for the grouped data can be calculated as

$$\widehat{X} = L_{mo} + w\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \tag{5}$$

where L_{mo} = the lower class boundary of modal class
w = the size of the modal class
$\Delta_1 = f_{mo} - f_1$
$\Delta_2 = f_{mo} - f_2$
f_{mo} = frequency of the modal class,
f_1= frequency of the class preceeding the modal class and
f_2 = frequency of the class following the modal class

Note: The modal class is a class with the highest frequency.

Example: Following is the distribution of the size of certain farms selected at random from a district. Calculate the mode of the distribution.

Size of farm	No of Farms
4.5-14.5	8
14.5-24.5	12
24.5-34.5	17
34.5-44.5	29
44.5-54.5	31
54.5-64.5	5
64.5-74.5	3

Solution:

Example: Following is the distribution of the size of certain farms selected at random from a district. Calculate the mode of the distribution.

Size of farm	No of Farms
4.5-14.5	8
14.5-24.5	12
24.5-34.5	17
34.5-44.5	29
44.5-54.5	31
54.5-64.5	5
64.5-74.5	3

Solution:44.5-54.5 is model class,$\Rightarrow L_{mo} = 44.5$
Frequency of the model class $f_{mo} = 31$, $w = 10$, $f_1 = 29$, $f_2 = 5$
$\Delta_1 = f_{mo} - f_1 \Rightarrow 31 - 29 = 2$, and
$\Delta_2 = f_{mo} - f_2 \Rightarrow 31 - 5 = 26$

Example: Following is the distribution of the size of certain farms selected at random from a district. Calculate the mode of the distribution.

Size of farm	No of Farms
4.5-14.5	8
14.5-24.5	12
24.5-34.5	17
34.5-44.5	29
44.5-54.5	31
54.5-64.5	5
64.5-74.5	3

Solution:44.5-54.5 is model class,$\Rightarrow L_{mo} = 44.5$
Frequency of the model class $f_{mo} = 31$, $w = 10$, $f_1 = 29$, $f_2 = 5$
$\Delta_1 = f_{mo} - f_1 \Rightarrow 31 - 29 = 2$, and
$\Delta_2 = f_{mo} - f_2 \Rightarrow 31 - 5 = 26$

$$\Rightarrow \widehat{X} = 45.5 + 10\left(\frac{2}{2+26}\right) = 46.214$$

Properties of mode

It is easy to locate , even by inspection in some data contexts

It can be found using graph (frequency curve)

It is not affected by extreme values

It can be calculated for distributions with open end classes

Often its value is not unique

Not calculated based on all the observations

The main drawback of mode is that often it may not exist

Which measure of central tendency is best with a given set of data?

Two factors are important in making this decisions:

Which measure of central tendency is best with a given set of data?

Two factors are important in making this decisions:

- The scale of measurement (type of data)
  The mean can be used for discrete and continuous data.

  The median is appropriate for discrete and continuous data as well, but can also be used for ordinal data

  The mode can be used for all types of data, but may be especially useful for nominal and ordinal measurements

Which measure of central tendency is best with a given set of data?

Two factors are important in making this decisions:

I The scale of measurement (type of data)
The mean can be used for discrete and continuous data.

The median is appropriate for discrete and continuous data as well, but can also be used for ordinal data

The mode can be used for all types of data, but may be especially useful for nominal and ordinal measurements

II The shape of the distribution of the observations

Symmetric and unimodal distribution: Mean, median, and mode should all be approximately the same

Which measure of central tendency is best with a given set of data?

Two factors are important in making this decisions:

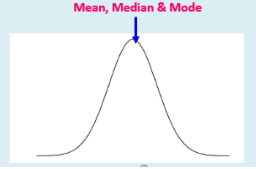
I The scale of measurement (type of data)
The mean can be used for discrete and continuous data.

The median is appropriate for discrete and continuous data as well, but can also be used for ordinal data

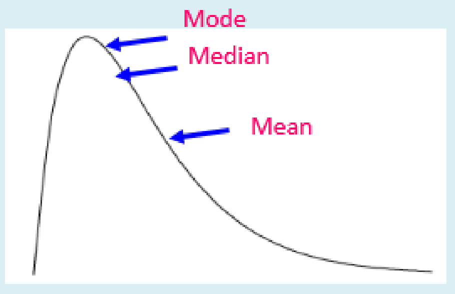
The mode can be used for all types of data, but may be especially useful for nominal and ordinal measurements

II The shape of the distribution of the observations

Symmetric and unimodal distribution: Mean, median, and mode should all be approximately the same



Skewed to the right (positively skewed) —Mean is sensitive to extreme values, so median might be more appropriate



The mean and median of symmetric distribution coincide.

When the distribution is skewed to the right, its mean is larger than its median.

When the distribution is skewed to the left, its mean is smaller than its median.

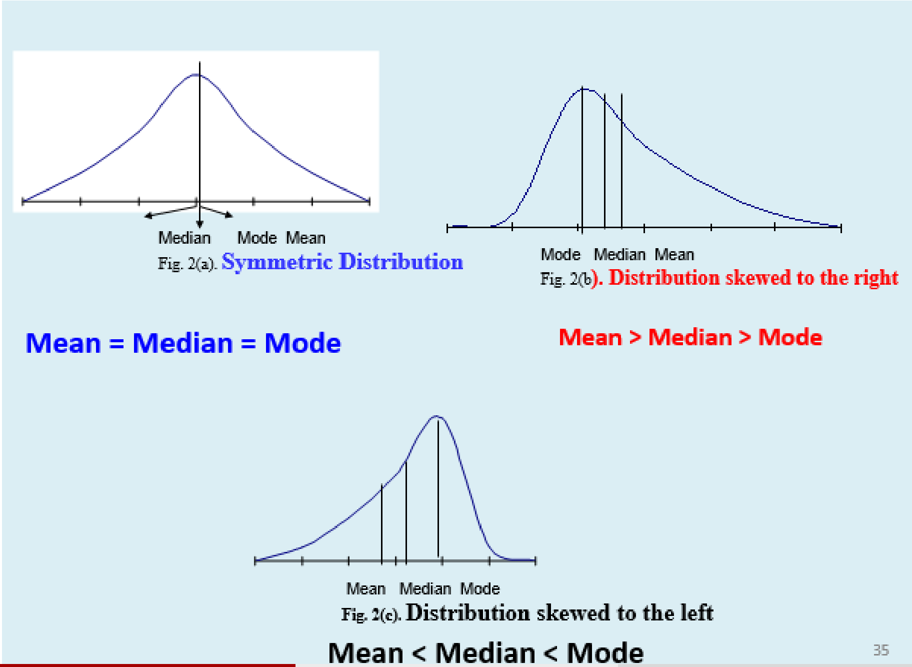


Fig. 2(a). Symmetric Distribution

Fig. 2(b). Distribution skewed to the right

Mean = Median = Mode

Mean > Median > Mode

Fig. 2(c). Distribution skewed to the left

Mean < Median < Mode

35

Reading Assignment

Geometric Mean

Harmonic Mean

Weighted Mean

Percentiles, Quartiles, Deciles

Measures of Dispersion

The scatter or spread of items of a distribution is known as dispersion or variation.

The degree to which numerical data tend to spread or scattered about an average value is called measure of variation or dispersion of the data.

The amount may be small when the values are close together.

If all the values are the same, no dispersion

Objectives of measuring Variation

Measures of Dispersion

The scatter or spread of items of a distribution is known as dispersion or variation.

The degree to which numerical data tend to spread or scattered about an average value is called measure of variation or dispersion of the data.

The amount may be small when the values are close together.

If all the values are the same, no dispersion

Objectives of measuring Variation

To judge the reliability of measures of central tendency

To control variability itself.

To compare two or more groups of numbers in terms of their variability.

To make further statistical analysis.

Absolute and Relative Measures of Dispersion

The measures of dispersion which are expressed in either absolute or relative terms

An absolute measure of variation is a measure which is expressed in terms of the unit that the original data is given

Such measures are not suitable for comparing the variability of two distributions which are expressed in different units of measurement and different average size

Relative measures of dispersions are a dimensionless quantity

It is obtained by dividing the absolute measure of variation to an appropriate measure of central tendency

Absolute measure of variation can be used for comparing variation in two groups only if the variable in those groups are measured in the same unit of measurement.

The most commonly used measures of dispersions are:

The most commonly used measures of dispersions are: Range, quartile deviation, coefficient of quartile deviation, Mean deviation, coefficient of mean deviation, Variance, standard deviation, coefficient of variation and standard scores

The most commonly used measures of dispersions are: Range, quartile deviation, coefficient of quartile deviation, Mean deviation, coefficient of mean deviation, Variance, standard deviation, coefficient of variation and standard scores

Range (R)

It is the simplest and the crudest measure of variation.

The range is defined as the difference between the highest and smallest obeservation in the data.

Range = Largest observation - Smallest observation (L-S)

The most commonly used measures of dispersions are: Range, quartile deviation, coefficient of quartile deviation, Mean deviation, coefficient of mean deviation, Variance, standard deviation, coefficient of variation and standard scores

Range (R)

It is the simplest and the crudest measure of variation.

The range is defined as the difference between the highest and smallest obeservation in the data.

Range = Largest observation - Smallest observation (L-S)

Example: consider the following data

60, 40, 30, 50, 60, 40, 70

Range= 70-30=40

properties of Range

It is the simplest crude measure of variation

It is not based on all observation

It is highly unstable measure

Does not give any information about the characteristic of the observation

It takes into account only two values which causes it to be a poor measure of dispersion.

Very sensitive to extreme observation

Quartiles and Inter-quartile Range

Percentiles divide the data into 100 parts of observations in each part.

It follows that the 25^{th} percentile is the first quartile, the 50^{th} percentile is the median and the 75^{th} percentile is the third quartile.

The quartiles are sets of values which divide the distribution into four parts such that there are an equal number of observations in each part.

$$Q_1 = [(n+1)/4]^{th}$$
$$Q_2 = [2(n+1)/4]^{th}$$
$$Q_3 = [3(n+1)/4]^{th}$$

The inter-quartile range(IQR) is the difference between the third and the first quartiles. $IQR = Q_3 - Q_1$

It is not affected by extreme values

The quartile deviation is given by

$$Q.D = \frac{Q_3 - Q_1}{2}$$

Example 1: find the IQR for the following data.

18, 21, 23, 24, 24, 32, 42, 59, 67,68,70

Since it is already arranged data

Example 1: find the IQR for the following data.
18, 21, 23, 24, 24, 32, 42, 59, 67,68,70
Since it is already arranged data

Calculate the values of $(n+1)/4$ and $3(n+1)/4$ as 12/4 and 36/4 which is equal to 3 and 9 respectively.

The observations lying in the 1^{st} position is 23 and 3^{rd} position is 67.

$\text{IQR} = (Q_3 - Q_1) = (67 - 23) = 44$

Example 2: find the IQR for the following data.
12, 15, 17, 25, 27,29, 34, 37

Example 1: find the IQR for the following data.
18, 21, 23, 24, 24, 32, 42, 59, 67,68,70
Since it is already arranged data

Calculate the values of $(n+1)/4$ and $3(n+1)/4$ as 12/4 and 36/4 which is equal to 3 and 9 respectively.

The observations lying in the 1^{st} position is 23 and 3^{rd} position is 67.

$\text{IQR} = (Q_3 - Q_1) = (67 - 23) = 44$

Example 2: find the IQR for the following data.
12, 15, 17, 25, 27,29, 34, 37

Then, (n+1)/4 = 2.25 and 3(n+1)/4 = 6.75

$Q_1 = 2.25^{th}$ observation= 15+(17-15)*0.25 = 15.5

$Q_3 = 6.75^{th}$ observation = 29+(34-29)*0.75 = 32.75

IQR=32.75-15.5=17.25

Variance

The average of the squared deviation from the mean.

Population variance measuring denoted by σ^2 (“sigma-squared”).

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

Measuring variance of sample denoted by s^2 (“s-squared”).

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

The main objection of mean deviation, that the negative signs are ignored, is removed by taking the square of the deviations from the mean

It is squared because the sum of the deviations of the individual observations of a sample about the sample mean is always 0

The variance can be thought of as an average of squared deviations.

Variance is used to measure the dispersion of values relative to the mean

When values are close to their mean (narrow range) the dispersion is less

Computation of Variance

i) For ungrouped data

Variance is used to measure the dispersion of values relative to the mean

When values are close to their mean (narrow range) the dispersion is less

Computation of Variance

i) For ungrouped data

Let $x_1, x_2, ..., x_n$ be n observation on a given data, the

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{k} x_i^2 - n\overline{x}^2}{n-1}$$

If x_i occurs f_i times for i=1, 2, ..., k then

$$s^2 = \frac{\sum_{i=1}^{n} f_i(x_i - \overline{x})^2}{\sum_{i=1}^{n} f_i - 1} = \frac{\sum_{i=1}^{n} f_i x_i^2 - n\overline{x}^2}{\sum_{i=1}^{n} f_i - 1}$$

ii) For grouped data

Let $x_1, x_2, ..., x_k$ be the class mark of grouped frequency with its corresponding frequency $f_1, f_2, ..., f_n$ respectively, then

ii) For grouped data

Let $x_1, x_2, ..., x_k$ be the class mark of grouped frequency with its corresponding frequency $f_1, f_2, ..., f_n$ respectively, then

$$s^2 = \frac{\sum_{i=1}^{n} f_i(x_i - \overline{x})^2}{n-1}$$

where $n = \sum f_i$ and $\overline{x} = \frac{\sum^k f_i x_i}{n}$

Properties of Variance

ii) For grouped data

Let $x_1, x_2, ..., x_k$ be the class mark of grouped frequency with its corresponding frequency $f_1, f_2, ..., f_n$ respectively, then

$$s^2 = \frac{\sum_{i=1}^{n} f_i(x_i - \overline{x})^2}{n-1}$$

where $n = \sum f_i$ and $\overline{x} = \frac{\sum^k f_i x_i}{n}$

Properties of Variance

The main disadvantage of variance is that its unit is the square of the unite of the original measurement values.

The variance gives more weight to the extreme values as compared to those which are near to mean value, because the difference is squared in variance.

The drawbacks of variance are overcome by the standard deviation.

Standard deviation

Sample standard deviation is square root of sample variance, and so is denoted by S.

Units are the original units.

Measures average deviation of data points from their mean.

Also, highly affected by outliers.

Calculation of SD in case of Raw data

$SD = \sqrt{variance} = S$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}f_i(x_i - \overline{x})^2}$$

Properties of the Standard Deviation

The standard deviation has the same units of measurement as the variable.

If the observations are expressed in centimeters, the standard deviation is expressed in centimeters.

If a constant value is added to each of the observations, the value of the standard deviation is unchanged.

If the observations are multiplied by a positive constant value, the standard deviation is multiplied by the same constant value.

In many cases it is true that approximately 68% of the observations fall within one standard deviation of the mean; approximately 95% within two standard deviations.

However, if the units of measurements of variables of two data sets is not the same, then there variability can't be compared by comparing the values of SD.

Coefficient of Variation

It is defined as the ratio of standard deviation to the mean values and is usually expressed in percentage

$$C.V = \frac{s}{\bar{x}} * 100\%$$

for the sample

$$C.V = \frac{\sigma}{\mu} * 100\%$$

population parameter

Note C.V is one of the most widely used measure of variation.

When two data sets have different units of measurements, or their means differ sufficiently in size, the CV should be used as a measure of dispersion.

It is the best measure to compare the variability of two series of sets of observations.

Data with less coefficient of variation is considered more consistent.

Example: the following summary statistics on two groups say A,B

	Mean	Median	Variance
Group A	52.5	50.5	100
Group B	47.5	45.5	121

In which of the group is the greater variable?

Chapter Four

4. Probability theory and probability distribution

Learning Objectives

Understand the concepts and characteristics of probabilities and probability distributions

Compute probabilities of events and conditional probabilities

Differentiate between the binomial and normal distributions

Understand the concepts and uses of the standard normal distribution

Probability theory and probability distribution

- Suppose it is known that a specific treatment is effective in 70% of the patients receiving the treatment
- This implies that the population consists of patients for whom the treatment is not effective about (30%) as well as patients for whom the treatment does have an effect is (70%)
- If the treatment is administered to 100 randomly chosen patients, more than 70 may experience improvement, or less than 70
- Question: If 100 patients are given the treatment, what is the probability that less than 60 of them will experience an improvement?
- Probability theory aims at predicting the out-come of an experiment, knowing the population
- These examples suggest the chance of an occurrence of some event of a random variable.

Why Probability in Medicine?

Because medicine is an inexact science, physicians seldom predict an outcome with absolute certainty.

E.g., to formulate a diagnosis, a physician must rely on available diagnostic information about a patient

History and physical examination

Laboratory investigation, X-ray findings, ECG, etc

An understanding of probability is fundamental for quantifying the uncertainty that is inherent in the decision-making process.

Probability theory also allows us to draw conclusions about a population based on known information about a sample which drown from that population

Definition

Probability is the chance of an outcome of an experiment. It is the measure of how likely an outcome is to occur.

Experiment: Any process of observation or measurement or any process which generates well defined outcome.

Random experiment/ Probability Experiment: It is an experiment that can be repeated any number of times under similar conditions and it is possible to enumerate the total number of outcomes with out predicting an individual out come.

Outcome :The result of a single trial of a random experiment

Sample Space: The set of all possible outcomes of an experiment that can be countable or uncountable.

Events: It is a subset of sample space.

Equally likely events: Events which have the same chance of occurring.

Mutually Exclusive Events (Disjoint Events) : Two events which cannot happen at the same time.

Equally likely events: Events which have the same chance of occurring.

Mutually Exclusive Events (Disjoint Events) : Two events which cannot happen at the same time.

Independent events The occurrence or non-occurrence of one event doesn't affect the occurrence or non-occurrence of the other event in repeated trials,

Equally likely events: Events which have the same chance of occurring.

Mutually Exclusive Events (Disjoint Events) : Two events which cannot happen at the same time.

Independent events The occurrence or non-occurrence of one event doesn't affect the occurrence or non-occurrence of the other event in repeated trials,

While tossing of two coin simultaneously, the occurrence of head in one coin does not affect the occurrence of tail on the other.

Counting Rules

In order to calculate probabilities, we have to know

The number of elements of an event

The number of elements of the sample space.

That is in order to judge what is probable, we have to know what is possible.

In order to determine the number of outcomes, one can use several rules of counting.

- Addition rule
- The multiplication rule
- Permutation rule
- Combination rule

The addition rule

If 1^{st} procedure designed by 1 can be performed in n_1 ways.

2^{nd} procedure designed by 2 can be performed in n_2 ways.

suppose further more that, it is not possible that both procedures 1 and 2 are performed together then the number of ways in which we can perform 1 or 2 procedure is $n_1 + n_2$ ways, and also if we have another procedure that is designed by k with possible way of n_k we can conclude that there is $n_1 + n_2 + \cdots + n_k$ possible ways.

Example: suppose we planning a trip and are deciding by bus and train transportation. If there are 3 bus routes and 2 train routes to go from A to B. find the available routes for the trip.

Solution: There are 3+2 =5 routes for someone to go from A to B.

The Multiplication rule

If a choice consists of k steps of which the first can be made in n_1 ways, the second can be made in n_2 ways $\ldots$, the k^{th} can be made in n_k ways, then the whole choice can be made in $n_1 * n_2 * \cdots * n_k$ ways

Example: Distribution of Blood Types There are four blood types, A, B, AB, and O.

Blood can also be Rh+ and Rh-. Finally, a blood donor can be classified as either male or female. How many different ways can a donor have his or her blood labeled?

Solution Since there are 4 possibilities for blood type, 2 possibilities for Rh factor, and 2 possibilities for the gender of the donor, there are 4 * 2* 2, or 16, different classification categories.

Permutation

An arrangement of n objects in a specified order

Permutation Rules:

1 The number of permutations of n distinct objects taken all together is n!

$$n! = n * (n-1) * (n-2) * \cdots * 2 * 1, \quad nP_n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!$$

2 The arrangement of n objects in a specified order using r objects at a time nP_r

3 The number of permutations of n objects in which k_1 are alike k_2 are alike . . . etc is

$$nP_k = \frac{n!}{k_1! * k_2 * \cdots * k_n}$$

Example: Suppose we have a letters A,B, C, D

a How many permutations are there taking all the four?

b How many permutations are there two letters at a time?

How many different permutations can be made from the letters in the word "CORRECTION"?

Combination rule

A selection of objects with out regard to order

Example:

1. In how many ways 5 patients be chosen out of 9 patients?
2. Out of 5 Mathematician and 7 Statistician a committee consisting of 2 Mathematician and 3 Statistician is to be formed. In how many ways this can be done if
   a. There is no restriction
   b. One particular Statistician should be included
   c. Two particular Mathematicians can not be included on the committee

Approaches to measuring Probability

There are four different conceptual approaches to the study of probability theory. These are:

1. Classical approach
2. Relative frequency approach
3. The axiomatic approach
4. The subjective approach.

Classical approach

This approach is used when all outcomes are equally likely

Definition: If a random experiment with N equally likely outcomes is conducted and out of these N_A outcomes are favorable to the event A, then the probability that event A occur denoted is defined as:

$$P(A) = N_A/N$$

Example: What is the probability of getting 6 when rolling a well-balanced die? An odd number? An even number?

Short coming of the classical approach

The total number of outcomes is infinite.

Outcomes are not equally likely.

Relative Frequency Probability

This is based on the relative frequencies of outcomes belonging to an event.

The probability of an event A is the proportion of outcomes favorable to A in the long run when the experiment is repeated under same condition

Example: If records show that 60 out of 100,000 bulbs produced are defective. What is the probability of a newly produced bulb to be defective?

Solution: Let A be the event that the newly produced bulb is defective

$$P(A) = lim_{n\rightarrow\infty}\frac{N_a}{N} = \frac{60}{100000} = 0.0006$$

Axiomatic Approach

Let E be a random experiment and S be a sample space associated with E.

With each event A a real number called the probability of A satisfies the following properties called axioms of probability or postulates of probability.

1. $P(A) \geq 0$
2. $P(S) = 1$, S is the sure event.
3. If A and B are mutually exclusive events, the probability that one or the other occur equals the sum of the two probabilities i.e $P(A \cup B) = P(A) + P(B)$
4. $P(A') = 1 - P(A)$
5. $0 \leq P(A) \leq 1$
6. $P(\emptyset) = 0, \quad \emptyset$ is the impossible event.

Conditional probabilities and the multiplicative law

Sometimes the chance a particular event happens depends on the outcome of some other event. This applies obviously with many events that are spread out in time.

Example: The chance a patient with some disease survives the next year depends on his having survived to the present time. Such probabilities are called conditional.

The notation is Pr(B/A), which is read as “the probability of occurrence of event B given that event A has already occurred .”

Let A and B be two events of a sample space S. The conditional probability of an event A, given B, denoted by

$$Pr(A/B) = \frac{P(A \cap B)}{P(B)}, \; P(B) \neq 0.$$

Independent Events

Two events are independent if the occurrence of one of the events does not affect the probability of the other event.

That is, A and B are independent if :

$$P(A \cap B) = P(A) \times P(B)$$

Hence

$$P(B|A) = P(B) \text{ or } P(A|B) = P(A)$$

Example: Let event A stands for “the sex of the first child from a mother is female”; and event B stands for “the sex of the second child from the same mother is male” Are A and B independent?

Solution P(B/A) = P(B) = 0.5 The occurrence of A does not affect the probability of B, so the events are independent

Multiplication rule

If A and B are independent events, then

$$P(A \cap B) = P(A) \times P(B)$$

$$P(A|B) = P(A), P(B) \neq 0$$

$$P(B|A) = P(B), P(A) \neq 0$$

P(A and B) denotes the probability that A and B both occur at the same time.

Example

Calculating probability of an event
Table 1: Shows the frequency of cocaine use by sex among adult cocaine users

Example

Calculating probability of an event
Table 1: Shows the frequency of cocaine use by sex among adult cocaine users

Life time frequency of cocaine use	Male	Female	Total
1-19 times	32	7	39
20-99 times	18	20	38
more than 100 times	25	9	34
Total	75	36	111

Questions

1. What is the probability of a person randomly picked is a male?
2. What is the probability of a person randomly picked uses cocaine more than 100 times?
3. what is the probability of getting male given that the selected person uses cocaine less than 20 times?
4. Given that the selected person is male, what is the probability of a person randomly picked uses cocaine more than 100 times?
5. Given that the person has used cocaine less than 100 times, what is the probability of being female?

Solution

1. P(m)=Total adult males/Total adult cocaine users =75/111 =0.68.

Solution

1. P(m)=Total adult males/Total adult cocaine users =75/111 =0.68.
2. $P(C > 100) = \frac{\text{All adult cocaine users more than 100 times}}{\text{Total adult cocaine users}}$

Solution

1. P(m)=Total adult males/Total adult cocaine users =75/111 =0.68.
2. $P(C > 100) = \frac{\text{All adult cocaine users more than 100 times}}{\text{Total adult cocaine users}}$

$$= \frac{34}{111} = 0.31.$$

3. $P(M|C < 20) = \frac{P(M \cap C < 20)}{P(C < 20)} = \frac{32/111}{39/111} = \frac{0.29}{0.35} = 0.83$

Solution

1. P(m)=Total adult males/Total adult cocaine users =75/111 =0.68.
2. $P(C > 100) = \frac{\text{All adult cocaine users more than 100 times}}{\text{Total adult cocaine users}}$

$$= \frac{34}{111} = 0.31.$$

3. $P(M|C < 20) = \frac{P(M \cap C < 20)}{P(C < 20)} = \frac{32/111}{39/111} = \frac{0.29}{0.35} = 0.83$
4. $P(C > 100|m) = \frac{P(C > 100 \cap m)}{P(m)} = \frac{25/111}{75/111} = 0.23/0.68 = 0.34$

Solution

1. P(m)=Total adult males/Total adult cocaine users =75/111 =0.68.
2. $P(C > 100) = \frac{\text{All adult cocaine users more than 100 times}}{\text{Total adult cocaine users}}$
$$= \frac{34}{111} = 0.31.$$
3. $P(M|C < 20) = \frac{P(M \cap C < 20)}{P(C < 20)} = \frac{32/111}{39/111} = \frac{0.29}{0.35} = 0.83$
4. $P(C > 100|m) = \frac{P(C > 100 \cap m)}{P(m)} = \frac{25/111}{75/111} = 0.23/0.68 = 0.34$
5. $P(f|C < 100) = \frac{P(f \cap C < 100)}{P(f)} = \frac{9/111}{34/111} = 0.081/0.306 = 0.265$

Solution

1. P(m)=Total adult males/Total adult cocaine users =75/111 =0.68.
2. $P(C > 100) = \frac{\text{All adult cocaine users more than 100 times}}{\text{Total adult cocaine users}}$
$$= \frac{34}{111} = 0.31.$$
3. $P(M|C < 20) = \frac{P(M \cap C < 20)}{P(C < 20)} = \frac{32/111}{39/111} = \frac{0.29}{0.35} = 0.83$
4. $P(C > 100|m) = \frac{P(C > 100 \cap m)}{P(m)} = \frac{25/111}{75/111} = 0.23/0.68 = 0.34$
5. $P(f|C < 100) = \frac{P(f \cap C < 100)}{P(f)} = \frac{9/111}{34/111} = 0.081/0.306 = 0.265$

Summary of basic Properties of probability

Probabilities are real numbers on the interval from 0 to 1; i.e.,

$$0 \leq P(A) \leq 1$$

If an event is certain to occur, its probability is 1, and if an event is certain not to occur, its probability is 0.

If two events are mutually exclusive (disjoint), the probability that one or the other will occur equals the sum of the probabilities;

$$P(AorB) = P(A) + P(B)$$

If A and B are two events, not necessarily disjoint, then

$$P(AorB) = P(A) + P(B) - P(AandB)$$

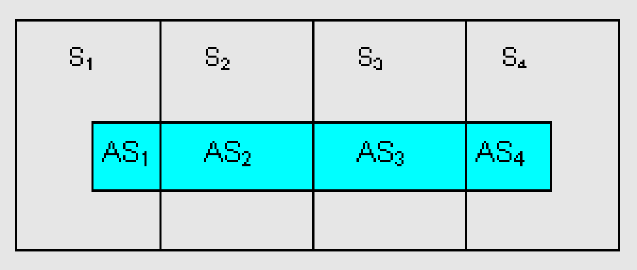
The sum of the probabilities that an event will occur and that it will not occur is equal to 1; i.e., $P(A') = 1 - P(A)$

Total probability rule

Assume the set of events $\{S_1, S_2, \ldots, S_n\}$ is a partition of a sample space S. Assume $P(S_i) > 0$ for every $i, 1 \leq i \leq n$.

Then for any event A,

$$P(A) = \sum_{i=1}^{n} P(S_i).P(A/S_i)$$



Bayes' rule

Assume the set of events $\{S_1, S_2, \ldots, S_n\}$ is a partition of a sample space S. Assume $P(S_i) > 0$ for every $i, 1 \leq i \leq n$.

Fix any event A.

$$P(S_j|A) = \frac{P(S_i).P(A/S_i)}{\sum_{i=1}^{n} P(S_i).P(A/S_i)}$$

Example: An insurance company issues life insurance policies in three separate categories: standard, preferred, and ultra- preferred. Of the company's policyholders, 50% are standard,40% are preferred, and 10% are ultra-preferred. Each standard policyholder has probability 0 .010 of dying in the next year, each preferred policyholder has probability 0 .005 of dying in the next year, and e ach ultra-preferred policyholder has probability 0 .001 of dying in the next year. A policyholder dies in the next year.

a. What is the probability of "death in the next year"?

b. What is the probability that the deceased policyholder was ultra-preferred?

Solution: Let S , P , U denote the standard, preferred, and ultra-preferred policyholders, and le t D denote the event “dies in the next year”.

Given $P(U) = 0.1, \quad P(P) = 0.4, \quad P(S) = 0.5$
$P(D|U) = 0.001, \quad P(D|P) = 0.005, \quad P(D|S) = 0.01$

a. $P(D) = P(D/U)P(U) + P(D|P)P(P) + P(D|S)P(S)$
$= 0.1 * 0.001 + 0.005 * 0.4 + 0.01 * 0.5 = 0.0071$

b. $P(U|D) = \frac{P(U \cap D)}{P(D)} = \frac{P(D|U)P(U)}{P(D/U)P(U)+P(D|P)P(P)+P(D|S)P(S)}$

$$= \frac{0.001 * 0.1}{0.1 * 0.001 + 0.005 * 0.4 + 0.01 * 0.5} = \frac{0.0001}{0.0071} = 0.01408$$

Example: Up on arrival at a hospital's emergency room, patients are categorized ac cording to their condition as critical, serious , or stable . In the past year:

i 10% of the emergency room patients were critical;

ii 30% of the emergency room patients were serious;

iii the rest of the emergency room patients were s table;

iv 40% of the critical patients died;

v 10% of the serious patients died; and

vi 1% of the stable patients died.

Given that a patient survived, what is the probability that the patient was categorized as serious up on arrival?

Random variables and probability distributions

A random variable is a numerical description of the outcomes of the experiment or a numerical valued function defined on sample space, usually denoted by capital letters.

Random variables and probability distributions

A random variable is a numerical description of the outcomes of the experiment or a numerical valued function defined on sample space, usually denoted by capital letters.

Example: If X is a random variable, then it is a function from the elements of the sample space to the set of real numbers. i.e.

X is a function $X : S \to R$

Usually numbers can be associated with the outcomes of an experiment.

A random variable takes a possible outcome and assigns a number to it.

For example, the number of heads that come up when a coin is tossed four times is 0, 1,2,3 or 4. Sometimes, we may find a situation where the elements of a sample space are categories.

Random variables are of two types

- Discrete random variable: are variables which can assume only a specific number of values. They have values that can be counted

Random variables are of two types

- Discrete random variable: are variables which can assume only a specific number of values. They have values that can be counted

  Number of children in a family.
  Number of car accidents per week.
  Number of defective items in a given company.
  Number of bacteria per two cubic centimeter of water
- Continuous random variable: are variables that can assume all values between any two given values.

Random variables are of two types

- Discrete random variable: are variables which can assume only a specific number of values. They have values that can be counted

  Number of children in a family.
  Number of car accidents per week.
  Number of defective items in a given company.
  Number of bacteria per two cubic centimeter of water

- Continuous random variable: are variables that can assume all values between any two given values.

  weight of patients at hospital.
  Mark of a student.
  Life time of light bulbs.
  Length of time required to complete a given training

Probability Distribution

A probability distribution consists of a value of a random variable can assume and the corresponding probabilities of the values.

It is the way data are distributed, in order to draw conclusions about a set of data

The values taken by a discrete random variable and its associated probabilities can be expressed by a rule, or relationship that is called a probability mass (density) function.

Properties of Probability Distribution

Probability Distribution

A probability distribution consists of a value of a random variable can assume and the corresponding probabilities of the values.

It is the way data are distributed, in order to draw conclusions about a set of data

The values taken by a discrete random variable and its associated probabilities can be expressed by a rule, or relationship that is called a probability mass (density) function.

Properties of Probability Distribution

1. Since the values of a probability distribution are probabilities, they must be numbers in the interval from 0 to 1.
2. Since a random variable has to take on one of its values, the sum of all the values of a probability distribution must be equal to 1

$$\begin{aligned} 1. \qquad P(x) &\geq 0, \quad \text{if X is discrete} \\ f(x) &\geq 0, \quad \text{if X is continuous} \end{aligned}$$

$$1. \qquad P(x) \geq 0, \quad \text{if X is discrete}$$

$$f(x) \geq 0, \quad \text{if X is continuous}$$

$$2. \quad \sum_x P(X = x) = 1, \quad \text{if X is discrete}$$

$$\int_x f(x)d(x) = 1, \quad \text{if X is continuous}$$

$$1. \qquad P(x) \geq 0, \quad \text{if X is discrete}$$

$$f(x) \geq 0, \quad \text{if X is continuous}$$

$$2. \quad \sum_{x} P(X = x) = 1, \quad \text{if X is discrete}$$

$$\int_{x} f(x)d(x) = 1, \quad \text{if X is continuous}$$

Note:If X is discrete random variable the

$$P(\alpha \leq X < b) = \sum_{x=a}^{b-1} P(X)$$

If X is a continuous random variable then

$$P(a < X < b) = \int_{a}^{b} f(x)dx$$

$$1. \quad P(x) \geq 0, \quad \text{if X is discrete}$$

$$f(x) \geq 0, \quad \text{if X is continuous}$$

$$2. \quad \sum_{x} P(X = x) = 1, \quad \text{if X is discrete}$$

$$\int_{x} f(x)d(x) = 1, \quad \text{if X is continuous}$$

Note:If X is discrete random variable the

$$P(\alpha \leq X < b) = \sum_{x=a}^{b-1} P(X)$$

If X is a continuous random variable then

$$P(a < X < b) = \int_{a}^{b} f(x)dx$$

Probability of a fixed value of a continuous random variable is zero.

$$1. \qquad P(x) \geq 0, \quad \text{if X is discrete}$$
$$f(x) \geq 0, \quad \text{if X is continuous}$$

$$2. \quad \sum_{x} P(X = x) = 1, \quad \text{if X is discrete}$$
$$\int_{x} f(x)d(x) = 1, \quad \text{if X is continuous}$$

Note:If X is discrete random variable the

$$P(\alpha \leq X < b) = \sum_{x=a}^{b-1} P(X)$$

If X is a continuous random variable then

$$P(a < X < b) = \int_{a}^{b} f(x)dx$$

Probability of a fixed value of a continuous random variable is zero.

Probability means area for continuous random variable.

Introduction to expectation

Definition: Let a discrete random variable X assume the values $X_1, X_2, \ldots X_n$ with the probabilities $P(X_1), P(X_2), \ldots, P(X_n)$ respectively. Then the expected value of X ,denoted as E(X) is defined as:

$$E(X) = X_1 p(X_1) + X_2 P(X_2) + \ldots + X_n P(X_n) = \sum_{i=1}^{n} X_i P(X_i)$$

Let X be a continuous random variable assuming the values in the interval (a, b) such that $\int_a^b f(x)dx = 1$, then

$$E(x) = \int_a^b x f(x) dx$$

The variance of X is given by

$$\text{variance of x} = var(x) = E(x^2) - (E(x))^2$$

A. Discrete Probability Distributions

- A discrete probability distribution describes how likely it is to observe specific values for a discrete random variable.
- Suppose X is the random variable 'sickness absence'

$$X = \begin{cases} 1 & \text{absence due to ilness} \\ 0 & \text{otherwise} \end{cases}$$

- X can only take the values 0 and 1
- The probability distribution of X describes the probability of observing a 0 or a 1, respectively

The following data shows the number of diagnostic services a patient receives

x	0	1	2	3	4	5
P(X=x)	0.671	0.229	0.053	0.031	0.01	0.006

What is the probability that a patient receives exactly 3 diagnostic services?

$$P(X = 3) = 0.031$$

What is the probability that a patient receives at most one diagnostic service?

$$P(X \leq 1) = P(X = 0) + P(X = 1)$$

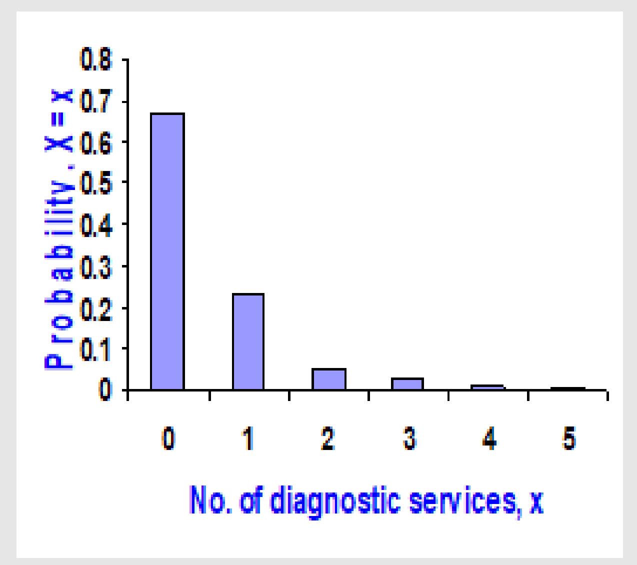
$$= 0.671 + 0.229 = 0.9$$

What is the probability that a patient receives at least four diagnostic services?

$$P(X \geq 4) = P(X = 4) + P(X = 5)$$

$$0.01 + 0.006 = 0.016$$

probability distribution can also be displayed using graph



- Many frequently used discrete distributions are given a name:
  - Bernoulli distribution
  - Multinomial distribution
  - Binomial distribution
  - Poisson distribution
  - Negative binomial distribution

Binomial Distribution

A binomial probability distribution occurs when the following requirements are met.

- Common probability distributions which is derived from a process known as a Bernoulli trial
- The procedure has a fixed number of trials n.
- The trials must be independent.
- Bernoulli trial is random process or experiment which can result in only one of two mutually exclusive outcomes (success or failure)
- The probability of a success(P) remains constant from trial to trial. for each trial

Binomial Distribution

A process that has only two possible outcomes is called a binomial process.

- Let $x_1, x_2 \cdots x_n$ denote outcomes for n independent and identical trials
- The probability distribution of outcome x for Y equals

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

  where $P(x_i = 1) = p$ and $p(x_i = 0) = 1 - p$
- The binomial distribution for $X = \Sigma_i x_i$ has mean and variance respectively
  - n denotes the number of fixed trials
  - **x** denotes the number of successes in the n trials
  - **p** denotes the probability of success
  - **q** denotes the probability of failure **(1-p)**

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Represents the number of ways of selecting x objects out of n where the order of selection does not matter.

where $n! = n(n-1)(n-2)\ldots(1)$ and $0! = 1$

Exercise

- Suppose that in a certain malarias area past experience indicates that the probability of a person with a high fever will be positive for malaria is 0.7. Consider 3 randomly selected patients (with high fever) in that same area.

a What is the probability that no patient will be positive for malaria?

b What is the probability that exactly one patient will be positive for malaria?

c What is the probability that exactly two of the patients will be positive for malaria?

d What is the probability that all patients will be positive for malaria?

Poisson Distribution

- The Poisson distribution is used for counts of events that occur randomly over time or space
- When some random events do not result from a fixed number of trials.
- if $x =$ number of deaths due to suicide accidents in Ethiopia during this coming week, there is no fixed upper limit n for x
- Since x must be a non-negative integer, its distribution should place its mass on that range
- A key feature of the Poisson distribution is that its variance equals its mean.
- Its probabilities depend on a single parameter, the mean λ

- The following are some examples which follow a Poisson process

- The following are some examples which follow a Poisson process
  - The number of telephone calls per hour at a switchboard
  - The number of e-mails received per hour
  - The number of patients admitted in a hospital emergency room per day
  - The number of defective items manufactured per 4-hour period in a manufacturing process. ETc

- The Poisson probability mass function is

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x = 0, 1, 2, \ldots$$

- Example: on the average 3-smokers pass a certain street corners every ten minutes, what is the probability that during a given 10 minutes the number of smokers passing will be

- The Poisson probability mass function is

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \qquad x = 0, 1, 2, \dots$$

- Example: on the average 3-smokers pass a certain street corners every ten minutes, what is the probability that during a given 10 minutes the number of smokers passing will be
  a. Exactly 5?
  b. at most 6?
  c. 7 or more?

Continuous probability distribution

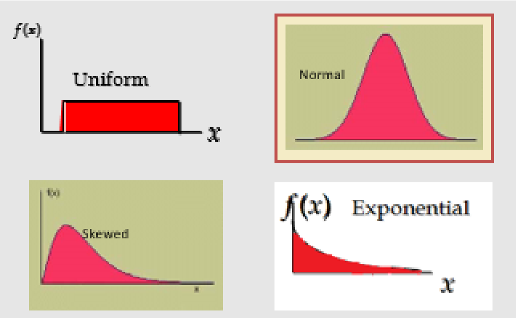
- A continuous probability distribution describes how likely it is that a continuous random variable takes values within certain ranges
- Some frequently used continuous distributions are

Continuous probability distribution

- A continuous probability distribution describes how likely it is that a continuous random variable takes values within certain ranges
- Some frequently used continuous distributions are
  - normal distribution,
  - chi-squared distribution
  - Exponential distribution

Continuous probability distribution

- A continuous probability distribution describes how likely it is that a continuous random variable takes values within certain ranges
- Some frequently used continuous distributions are
  - normal distribution,
  - chi-squared distribution
  - Exponential distribution



Normal Distribution

- The Normal Distribution is by far the most important probability distribution in statistics.
- The normal distribution is a theoretical, continuous probability distribution whose equation is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

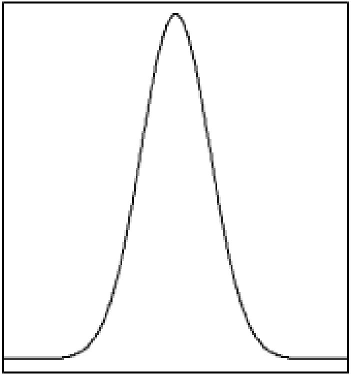
  where $-\infty < x < \infty$

- The two parameters of the normal distribution are the mean (μ) and the standard deviation (σ)
- The graph has a familiar bell-shaped curve.

Characteristics of the Normal Distribution

- It is a probability distribution of a continuous variable. It extends from minus infinity ($-\infty$) to plus infinity ($+\infty$)
- It is unimodal, bell-shaped and symmetrical about $x = \mu$
- The mean, the median and mode are all equal
- The total area under the curve above the x-axis is one square unit.
- The curve never touches the x-axis.
- It is determined by two quantities: its mean (μ) and SD (σ)
- An observation from a normal distribution can be related to a standard normal distribution (SND) which has a published table.

- The normal Distribution is a family of Bell-shaped and symmetric distributions as the allocation is symmetric: one-half (.50 or 50%) lies on either side of the mean



An observation from a normal distribution can be related to a standard normal distribution (SND) which has a published table.

Standard normal distribution

- Since the values of μ and σ will depend on the particular problem in hand and tables of the normal distribution cannot be published for all values of μ and σ calculations are made by referring to the standard normal distribution which has $\mu = 0$ and $\sigma = 1$.
- Thus an observation x from a normal distribution with mean μ and standard deviation σ can be related to a Standard normal distribution by calculating

$$SND = Z = \frac{(X - \mu)}{\sigma}$$

To find $P(a < x < b)$, we need to find the area under the appropriate normal curve.

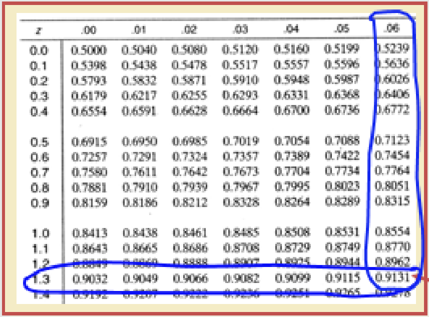
To simplify the tabulation of these areas, we standardize each value of x by expressing it as a z-score, the number of standard deviations s it lies from the mean m.

Properties of the Standard Normal Distribution

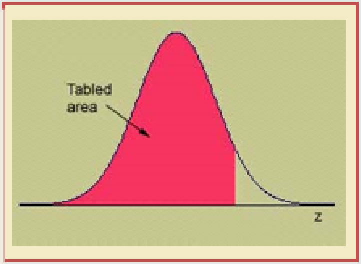
- Mean is zero
- Variance is one
- Standard Deviation is one
- Areas under the standard normal distribution curve have been tabulated in various ways.
- The most common ones are the areas between $Z = 0$ and a positive value of Z

The standard normal random variable, Z, is the normal random variable with mean $\mu = 0$ and standard deviation $\sigma = 1 : Z \sim N(0, 1^2)$

The four digit probability in a particular row and column of Table 1 gives the area under the z curve to the left that particular value of z.



z	.00	.01	.02	.03	.04	.05	.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051
0.9	0.8159	0.8186	0.8212	0.8328	0.8264	0.8289	0.8315
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279



Area for $z \leq 1.36$

- Given a normal distributed random variable X with Mean μ and standard deviation σ

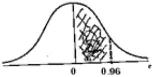
$$\begin{aligned} P(a < X < b) &= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \end{aligned}$$

Example: Find the area under the standard normal distribution which lies

- between $Z = 0$ and $Z = 0.96$
- between $Z = -1.45$ and $Z = 0$
- To the right of $Z = -0.35$
- Between $Z = -0.67$ and $Z = 0.75$

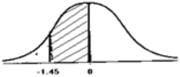
- Solution

a. $Area = P(0 < Z < 0.96) = 0.3315$



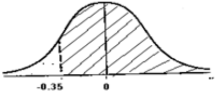
b.

$$\begin{aligned} Area &= P(-1.45 < Z < 0) \\ &= P(0 < Z < 1.45) \\ &= 0.4265 \end{aligned}$$



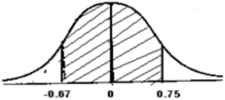
c.

$$\begin{aligned} Area &= P(Z > -0.35) \\ &= P(-0.35 < Z < 0) + P(Z > 0) \\ &= P(0 < Z < 0.35) + P(Z > 0) \\ &= 0.1368 + 0.50 = 0.6368 \end{aligned}$$



d.

$$\begin{aligned} Area &= P(-0.67 < Z < 0.75) \\ &= P(-0.67 < Z < 0) + P(0 < Z < 0.75) \\ &= P(0 < Z < 0.67) + P(0 < Z < 0.75) \\ &= 0.2486 + 0.2734 = 0.5220 \end{aligned}$$



- Example: A random variable X has a normal distribution with mean 80 and standard deviation 4.8. What is the probability that it will take a value
  - Less than 87.2
  - Greater than 76.4
  - Between 81.2 and 86.0
- Solution: X is normal with mean, $\mu = 80$, standard deviation, $\sigma = 4.8$

$$\begin{aligned} p(X < 87.2) &= P\left(\frac{X-\mu}{\sigma} < \frac{87.2-\mu}{\sigma}\right) \\ &= P\left(Z < \frac{87.2-80}{4.8}\right) \\ &= P(Z < 1.5) \\ &= P(Z < 0) + P(0 < Z < 1.5) \\ &= 0.5 + 0.4332 = 0.9332 \end{aligned}$$

Chapter Five

5. Sampling Method and Sample Size Calculation

Sampling method and sample size calculation

LEARNING OBJECTIVES

1. Define population and sample and understand the different sampling terminologies
2. Differentiate between probability and Non-Probability sampling methods and apply different techniques of sampling
3. Understand the importance of a representative sample
4. Differentiate between random error and bias
5. Enumerate advantages and limitations of the different sampling methods

Introduction

If we have to draw a sample, we will be confronted with the following questions:

- What is the group of people (population) from which we want to draw a sample?
- How many people do we need in our sample?
- How will these people be selected?
- What are the errors to be confronted with when taking a random sample?

Researchers often use sample survey methodology to obtain information about a larger population by selecting and measuring a sample from that population.

Since population is too large, for researchers as well as being studied in terms of time, money, privacy. . . ..feasibility

Inferences about the population are based on the information from the sample drawn for the population. If the whole population is taken, there is no need of statically inference.

However, due to the variability in the characteristics of the population, scientific sampling designs should be applied to select a representative sample.

If not, there is a high risk of distorting the view of the population.

Sampling is a procedure by which some members of the given population are selected as representative of the entire population.

Advantage of sampling

Reduced cost sampling reduces demands on resource such as finance, labour and material,

Greater speed data can be collected and summarized more quickly,

Quality data: due to The use of better trained personnel and more time and efforts can be spent on getting reliable data on each individual sampled.

It cover many important variables can not covered in DHS or censes

Disadvantage of Sampling

There is always sampling error

Sampling may create a feeling of discrimination within the population

It may be inadvisable where every unit in the population is legally required to have a record

Definition of terms used in sampling

Reference population (also called source population or target population): Is the population about which an investigator wishes to draw conclusion.

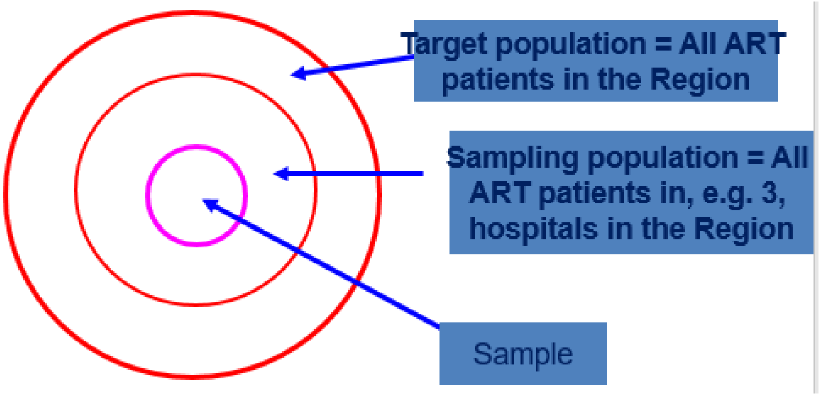
Study or sample population: Population from which the sample actually was drawn and about which a conclusion can be made. Or the subset of the target population from which a sample will be drawn.

Sampling unit : The unit of selection in the sampling process. For example sample of district, the sampling unite is district.

Study unit: The unit on which information is collected. if the objective is to determine the availability of latrine, then the study unit would be the household; if the objective is to determine the prevalence of trachoma, then the study unit would be the individual.

Sampling frame: The list of all the units in the reference population, from which a sample is to be picked.

Researchers are interested to know about factors associated with ART use among HIV/AIDS patients attending certain hospitals in a given Region



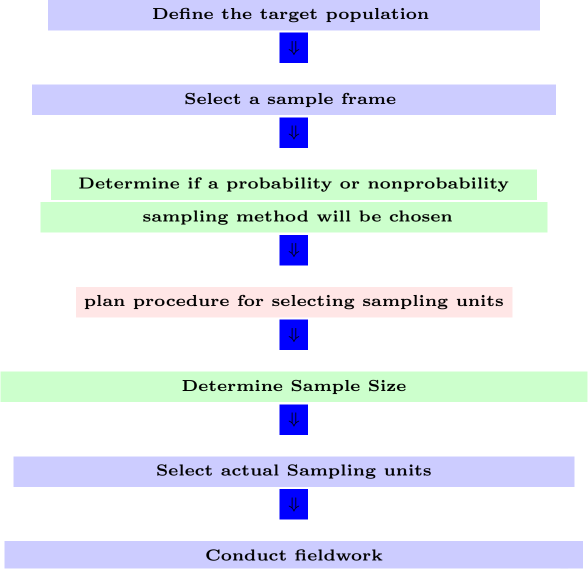
Target population = All ART patients in the Region

Sampling population = All ART patients in, e.g. 3, hospitals in the Region

Sample

Sampling is a process of selecting some members of a given population as representatives of the entire population in terms of the desired characteristics

Stages in selecting a sample



Define the target population

Select a sample frame

Determine if a probability or nonprobability sampling method will be chosen

plan procedure for selecting sampling units

Determine Sample Size

Select actual Sampling units

Conduct fieldwork

Sampling. . . why

Advantage

- Reduced cost
- Greater speed
- Greater scope
- Greater accuracy
- Feasibility

Disadvantage

- There is always sampling error
- Sampling may create a feeling of discrimination in the population.
- Inadvisable where every unit in the population is legally required to have a record

What are the error to be confronted with when taking a random sample?

When we take a sample, our results will not exactly equal the correct results for the whole population. That is , our result will be subject errors. Those errors has two components.

- Sampling errors(Random error)
- Non sampling error(bias)

Causes of sampling error

One is chance: That is the error that occurs just because of bad luck

Design error : Un representativeness of the sample

Sampling errors (random error) can be minimized by increasing the size of the sample. But can not void it completely .

Non-sampling error/BIAS

selection bias

- Design bias
- Accessibility bias
- Voluntary bias

Information Bias

- Observational error
- Respondent error/recall bias
- social desirability bias
- non response bias

It is possible to eliminated or reduce the non sampling Errors(bias) by careful design of the sampling procedure and by taking care of the errors that may be a rise during data analysis.

The best source of non sampling bias is non response.

It is failure to obtain information on some of the subjects include in the sample to be studied.

Non response bias is significant bias when the following two conditions are both fulfilled

- When non respondents constitute a significant proportion of the sample(about 15% and more)
- When non respondents differ significantly from respondents.

Errors in sampling

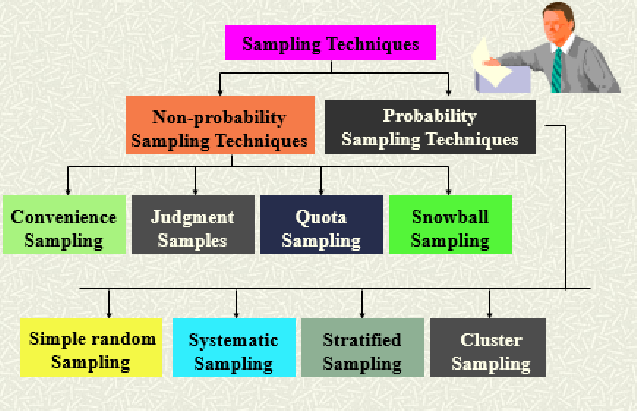
There are several ways to deal with this problem and reduce the possibility of bias:

- Data collection tools (questionnaire) have to be pre tested.
- If non response is due to absence of the subjects, repeated attempt should be considered to contact study subjects who were absent at the time of the initial visit.
- To include additional people in the sample, so that non respondents who were absent during data collection can be replaced (make sure that their absence is not related to the topic being studied).

N.B. : The number of non responses should be documented according to type, so as to facilitate an assessment of the extent of bias introduced by non response.

Sampling Methods/types

Two broad categories of sampling procedures:
probability methods and **non-probability methods**.



A. Probability sampling

Involves random selection of a sample

A sample is obtained in a way that ensures every member of the population to have a known (non zero) probability of being included in the sample.

Procedures to ensure that each unit of the sample is chosen on the basis of chance.

Every sampling unit has a known and non-zero probability of selection into the sample.

Sample finding can be generalizable

more complex, more time consuming and usually more costly

The method chosen depends on a number of factors, such as

The available sampling frame,

How spread out the population is,

How costly it is to survey members of the population

Homogeneity of the population

Most common probability sampling methods

1. Simple random sampling
2. Systematic random sampling
3. Stratified random sampling
4. Cluster sampling
5. Multi-stage sampling

1. Simple Random Sampling

The required number of individuals are selected at random from the sampling frame, a list or a database of all individuals in the population.

Each member of a population has an equal chance/probability of being included in the sample.

Representativeness of the sample is ensured

Procedure

- Take sampling population
- Make a numbered list of all the units in the population ("sampling frame")
- Each unit should be numbered from 1 to N (where N is the size of the population)
- Randomly draw the required numbers/units

The randomness of the sample is ensured by:

- Use of “lottery’ methods (sample drawn from box)
- Table of random numbers
- Computer generated random numbers

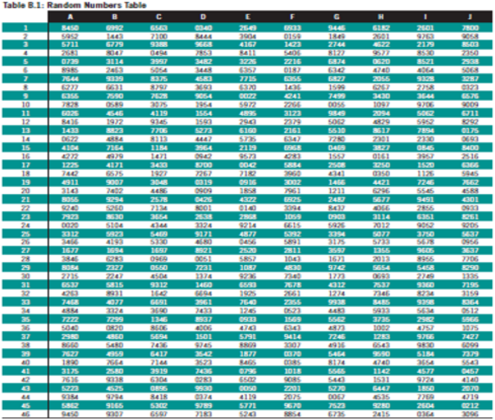


Table B.1: Random Numbers Table

	A	B	C	D	E	F	G	H	I	J
1	8450	6992	6563	0340	2649	6933	9446	6182	2601	7800
2	5952	1443	7100	8444	3904	0159	1849	2601	9763	9058
3	5711	6779	9388	9668	4167	1423	2744	4622	2179	8503
4	2683	8047	0494	7853	8411	5406	8127	9577	8530	2350
5	0739	3114	3997	3482	3226	2216	6874	0620	8521	2938
6	8985	2463	5054	3448	6357	0187	6342	4740	4064	5068
7	7644	9339	8375	4583	7715	6355	6827	2055	9328	3287
8	6277	6631	8797	3693	6370	1436	1599	6267	2758	0323
9	6355	7590	7628	9054	0022	4241	7499	3430	3644	6576
10	7828	0589	3075	1954	5972	2266	0055	1097	9706	9009
11	6026	4546	4119	1554	4895	3123	9849	2094	5062	6711
12	8416	1972	9345	1593	2943	2379	5062	4829	5952	8292
13	1433	8823	7706	5273	6160	2161	5510	8617	7894	0175
14	0622	4884	8113	4447	5735	6347	7280	2301	2330	0693
15	4104	7164	1184	3964	2119	6968	0469	3827	0845	8400
16	4272	4979	1471	0942	9573	4283	1557	0161	3957	2516
17	1225	4171	3433	8700	0042	5884	2508	3250	1520	6366
18	7442	6575	1927	7267	7182	3960	4341	0350	1126	5945
19	4911	9007	3048	0319	0916	3002	1466	4421	7246	7662
20	3143	7402	4486	0909	1858	7961	1211	6296	5545	4588
21	8055	9294	2578	0426	4322	6925	2487	5677	9491	4301
22	9240	5260	7134	8001	0140	3394	8437	4066	2856	0933
23	7923	8630	3654	2638	2868	1059	0903	3114	6351	8261
24	0020	5104	4344	3324	9214	6615	5926	7012	9052	9205
25	3312	5923	5469	9171	4877	5392	3394	5077	3750	5637
26	3466	4193	5330	4680	0456	5891	3175	5733	5678	0956
27	1677	1694	1697	8921	2520	2811	3597	1355	9605	3637
28	3846	6283	0969	0051	5857	1043	1671	2013	8956	7706
29	8084	2327	0550	7231	1087	4830	9742	5654	5458	8290
30	2715	2247	4504	1374	9236	7340	1773	0693	2749	1335
31	6537	5815	9312	1460	6593	7678	4312	7537	9360	7195
32	4263	8931	1642	6694	1925	2661	1274	7346	8234	3159
33	7468	4077	6691	3961	7640	2355	9938	8485	9398	8364
34	4884	3324	3690	7433	1245	0523	4483	5933	5634	0512
35	7222	7299	1346	8937	0933	1569	5562	3735	2982	5966
36	5040	0820	8606	4006	4743	6343	4873	1002	4757	1075
37	2980	4860	5694	1501	5791	9414	7246	1283	9766	7427
38	8660	5480	7436	9745	8869	3307	4916	6543	9830	6099
39	7627	4959	6417	3542	1877	0370	5464	9590	5184	7379
40	1890	7664	7144	3523	8465	0385	8174	4740	3654	5543
41	3175	2580	3919	7436	0796	1018	5565	1142	4577	0457
42	7616	9338	6304	0283	6502	9085	5443	1531	9724	4140
43	5223	4525	0896	9930	0050	2201	5270	6447	1850	2070
44	9384	9794	8418	0374	4119	2075	0067	4535	7769	4719
45	5862	9165	5302	9789	5771	9670	7523	9280	2604	0212
46	9450	9307	6597	7183	5243	8854	6735	2415	0364	3096

Advantages

- Simple
- Sampling error easily measured

Disadvantages

- Need complete list of units
- Units may be scattered and poorly accessible
- Heterogeneous population

2. Systematic Random Sampling

Sometimes called interval sampling

Selection of individuals from the sampling frame is done systematically rather than randomly.

Individuals are taken at regular intervals down the list (for example every k^{th})

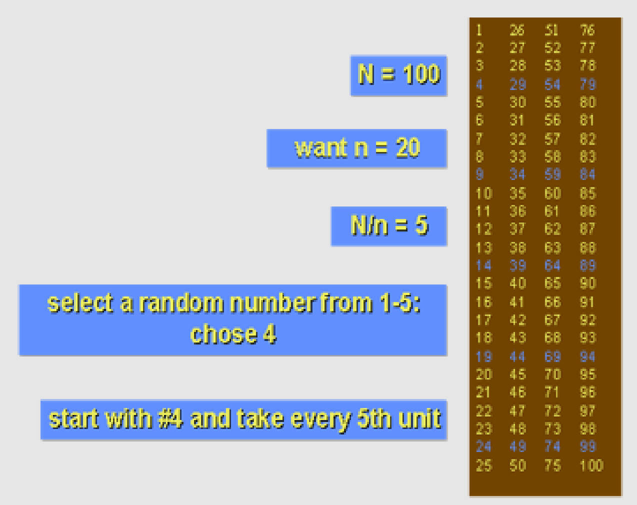
The first unit to be selected is taken at random from among the first K units.

Procedure

- Arrange the units in some kind of sequence (from 1 to N)
- Determine the sampling interval (K) by dividing the number of units in the population by the desired sample size (eg N/n=k)
- Choose a random starting point (for k, the starting point will be a random number between 1 and k)
- Select every k^{th} unit after that first number

To select a sample of 20 from a population of 100, you would need a sampling interval of 100 /20 = 5.

Therefore, K = 5.



N = 100

want n = 20

N/n = 5

select a random number from 1-5: chose 4

start with #4 and take every 5th unit

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

Advantage

Systematic sampling usually less time consuming and easier to perform than SRS. It provide good approximation to SRS

Unlike SRS, systematic sampling can be conducted without sampling frame (usually in some situation sampling frame not readily available)

Disadvantages

Periodicity-underlying pattern may be a problem (characteristics occurring at regular intervals)

3. Stratified random sampling

It is done when the population is known to have heterogeneity with regard to some factors and those factors are used for stratification

Using stratified sampling, the population is divided into homogeneous, mutually exclusive groups called strata, and

A population can be stratified by any variable that is available for all units prior to sampling (e.g., age, sex, province of residence, income, etc.).

A separate sample is taken independently from each stratum

Procedure

- Divide (stratify) sampling frame into homogeneous subgroups (strata) e.g. minorities, urban/rural areas, occupations
- Draw random sample within each stratum

The sampling method can vary from one stratum to another

Proportionate allocation- if the same sampling fraction is used for each stratum

Non-proportionate allocation- the strata unequal in size and a fixed number of units is selected from each stratum

Advantages

- representativeness of the sample is improved.
- focuses on important subpopulations and ignores irrelevant ones
- improves the accuracy of estimation

Disadvantages

- can be difficult to select relevant stratification variables
- not useful when there are no homogeneous subgroups
- can be expensive
- Sampling error is difficult to measure

Example: A sample of 50 students is to be drawn from a population consisting of 500 students belonging to two institutions A and B. The number of students in the institution A is 200 and the institution B is 300. How will you draw the sample using proportional allocation?

Solution: There are two strata in this case with sizes

$N_1 = 200$ and $N_2 = 300$ and the total population
$N = N_1 + N_2 = 500$ The sample size is 50.
If n_1 and n_2 are the sample sizes,

$$n_1 = \frac{n}{N} \times N_1 = \frac{50}{500} \times 200 = 20$$

$$n_2 = \frac{n}{N} \times N_2 = \frac{50}{500} \times 300 = 30$$

The sample sizes are 20 from A and 30 from B.

Then the units from each institution are to be selected by simple sampling

Advantage

The representativeness of the sample is improved . That is, adequate representation from each group.

Minority subgroups of interest can be ensured by stratification and by varying the sample fraction between strata as required.

Disadvantages

Sampling frame for the entire population has to be prepared separately for each stratum.

4. Cluster sampling

Sometimes it is too expensive to carry out SRS

- Population may be large and scattered.
- Complete list of the study population unavailable
- Travel costs can become expensive if interviewers have to survey people from one end of the country to the other.

Cluster sampling is the most widely used to reduce the cost

The clusters should be homogeneous, unlike stratified sampling where the strata are heterogeneous

Steps in cluster sampling

- Whole population divided into groups e.g. neighbourhoods
- A type of multi-stage sampling where all units at the lower level are included in the sample
- Random sample taken of these groups (“clusters”)

- Within selected clusters, all units e.g. households included (or random sample of these units)
- Provides logistical advantage

Involves selection of groups called clusters followed by selection of individuals within each selected cluster.

Can be used when it is either impossible or impractical to compile exhaustive list of individuals of the target population.

Cluster sampling is recommended for its efficiency, however accuracy is less because it is subject to more than one sampling error unlike SRS.

Advantages

- Simple as complete list of sampling units within population not required
- Less travel/resources required

Disadvantages

- Cluster members may be more alike than those in another cluster (homogeneous)
- this “dependence” needs to be taken into account in the sample size and in the analysis (“design effect”)

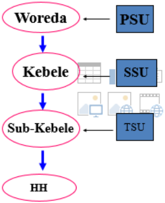
5. Multi-stage sampling

This method is appropriate when the reference population is large or widely scattered

The selection done in in stage until the final sampling unit(e.g. household, person) are arrived at.

The primary sampling unit (PSU) is the sampling unit in the first sampling stage.

The secondary sampling unit (SSU) is the sampling unit in the second sampling stage, etc.



In the first stage, large groups or clusters are identified and selected. These clusters contain more population units than are needed for the final sample.

In the second stage, population units are picked from within the selected clusters (using any of the possible probability sampling methods) for a final sample.

If more than two stages are used, the process of choosing population units within clusters continues until there is a final sample.

With multi-stage sampling, you still have the benefit of a more concentrated sample for cost reduction.

However, the sample is not as concentrated as other clusters and the sample size is still bigger than for a simple random sample size

Also, you do not need to have a list of all of the units in the population. All you need is a list of clusters and list of the units in the selected clusters.

Admittedly, more information is needed in this type of sample than what is required in cluster sampling.

However, multi-stage sampling still saves a great amount of time and effort by not having to create a list of all the units in a population.

B. Non-probability sampling

In non-probability sampling, every item has an unknown chance of being selected.

In non-probability sampling, there is an assumption that there is an even distribution of a characteristic of interest within the population.

For probability sampling, random is a feature of the selection process.

In non-probability sampling, since elements are chosen arbitrarily, there is no way to estimate the probability of any one element being included in the sample.

Also, no assurance is given that each item has a chance of being included, making it impossible either to estimate sampling variability.

Despite these drawbacks, non-probability sampling methods can be useful when descriptive comments about the sample itself are desired.

Secondly, they are quick, inexpensive and convenient.

The most common types of non-probability sampling

1. Convenience or haphazard sampling
2. Volunteer self selection sampling
3. Judgment/Purposive sampling
4. Quota sampling
5. Snowball sampling technique ... etc

1. Convenience or haphazard sampling

Convenience sampling is sometimes referred to as haphazard or accidental sampling.

It is not normally representative of the target population because sample units are only selected if they can be accessed easily and conveniently.

The obvious advantage is that the method is easy to use, but that advantage is greatly offset by the presence of bias.

Often used in face to face interviews

very easy to carry out,

Difficult to draw any meaningful conclusion. May not be representative

2. Volunteer sampling

As the term implies, this type of sampling occurs when people volunteer to be involved in the study.

Sampling voluntary participants as opposed to the general population may introduce strong biases.

Common in trials demanding long duration.

In psychological experiments or pharmaceutical trials (drug testing), for example, it would be difficult and unethical to enlist random participants from the general public.

Payments for subjects some times be involved.

3. Judgment sampling

This approach is used when a sample is taken based on certain judgments about the overall population.

The underlying assumption is that the investigator will select units that are characteristic of the population.

The critical issue here is objectivity: how much can judgment be relied upon to arrive at a typical sample?

Judgment sampling is subject to the researcher's biases and is perhaps even more biased than haphazard sampling.

Since any preconceptions the researcher may reflected in the sample, large biases can be introduced if these preconceptions are inaccurate.

One advantage of judgment sampling is the reduced cost and time involved in acquiring the sample.

4. Quota sampling

This is one of the most common forms of non-probability sampling.

Sampling is done until a specific number of units (quotas) for various sub-populations have been selected.

Since there are no rules as to how these quotas are to be filled, quota sampling is really a means for satisfying sample size objectives for certain sub-populations.

5. Snowball sampling

A technique for selecting a research sample where existing study subjects recruit future subjects among their friends.

Thus the sample group appears to grow like a rolling snowball.

This sampling technique is often used in hidden populations which are difficult for researchers to access; example populations would be drug users or commercial sex workers.

Because sample members are not selected from a sampling frame, snowball samples are subject to numerous biases. For example, people who have many friends are more likely to be recruited into the sample.

Sample size determination

In planning of any investigation we must decide how many people need to be studied to answer the study objectives .

If the study is too small, we may fail to detect important effects or if the study is too large we will waste resources.

In general, it is must better increase the accuracy of the data collection(by improving the training of data collection and data collection tools) than to increase sample size after a certain point.

That is called the minimal sample size required.

In order to calculate the required sample size, you need to know the following fact

The reasonable estimate of the key proportion to be studied. If you can not get, guess the proportion, take it as it has 50%.

The degree of accuracy required. This is the deviation from the true proportion in the population as the whole. It take with in 1% to 5%

The confidence level required usually specified as 95% at $\alpha = 0.05$ significance level.

The size of the population that the sample is to represent (N).

The minimum sample size required, for a very large population $(N > 10,000)$ is:

$$n_0 = \frac{Z^2_{\alpha/2} p(1-p)}{d^2}$$

If the researcher wants to include the size of the entire population (N) in the study, the above formula should be corrected using the correction formula:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Though it has no such significance difference, some scholars like (Israel, G.D. (1992) Sampling the Evidence of Extension Program Impact. Program Evaluation and Organizational Development, IFAS, University of Florida.) suggests a correction formula:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

Example 1

A study planned to find the determinant factors for the adherence and non adherence of HIV/ AIDS patients. From related literature it was found that 74% of the samples were adhere in using the ART with 95% confidence interval and 0.03 margin of error:

a) Find the sample size needed to carry out this research. Use $\alpha = 0.05$ and d=0.03.

Given: p = 0.74, d = 0.03, Z = 1.96 (i.e., for a 95% C.I.)

$$n_0 = \frac{Z^2_{\alpha/2} p(1-p)}{d^2}$$

$$= \frac{1.96^2(0.74 \times 0.26)}{0.03^2} = 821.25 \approx 822$$

Thus, the study should include at least 822 subjects.

b) If the above sample is to be taken from a relatively small population (say N = 3000), the required minimum sample will be obtained from the above estimate by making some adjustment.

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{821.25}{(1 + \frac{821.25}{3000})} \approx 645 \text{ subjects}$$

Exercise 2: A hospital administrator wishes to know what proportion of discharged patients is satisfied with the care received during hospitalization. From pilot study it was found that 70% of the patients satisfied with the service. With 95% CI and 4% accepted margin of error, how large a sample should be drawn to carry out this study? ($\approx$ 505).

Example 3

A hospital administrator wish to know what proportion of discharge patients are unhappy with the care received during hospitalization. If 95% confidence interval is desired to estimate the proportion within 5%, how large a sample should be drawn?

$$n_0 = \frac{Z^2_{\alpha/2} p(1-p)}{d^2}$$

$$= \frac{1.96^2(0.5 * 0.5)}{0.05^2} = 385$$

If you do not have any information about P , take it 50% and get the maximum value pq= 0.25

Rules of thumb

1. If the population size is small ($N < 100$), there is little point in sampling. Survey the entire population.
2. If the population size is around 500, 50% should be sampled.
3. If the population size is around 1500, 20% should be sampled.
4. Beyond a certain point (N=5000), the population size is almost irrelevant and sample size of 400 may be adequate.
5. Statistians maximalist at least 500

Chapter Six

6. Basic Concept of Inference

Basic Concept of Inference

Introduction to Statistical Inference

Statistical inference is a procedure whereby inferences about a population are made on the basis of the results obtained from a sample drawn from that population.

The term statistical inference deals with the collection of data on a relatively small number of cases so as to form conclusions about the general population from which the sample was taken.

Statistical Estimation

It is the procedure of using a sample statistics to estimate a population parameter.

A statistic used to estimate a parameter is called an estimator. And the value taken by the estimator is called an estimate.

Statistical estimation is divided in to two main categories: point estimate and Interval estimate.

Estimation is the process of determining a likely value for a variable in the survey population, based on information collected from the sample.

Estimation is the use of sample statistics to estimate unknown population parameters.

- A **sample statistic** is a numerical measure of a summary characteristic of a sample.

- A **population parameter** is a numerical measure of a summary characteristic of a population.

- An **estimator** of a population parameter is a sample statistic used to estimate or predict the population parameter.
- An **estimate** of a parameter is a particular numerical value that obtain using an estimator.
- A **point estimate** is a single value used as an estimate of a population parameter.

- A **Interval Estimation** the interval of values as an estimate for a parameter, which is interval that contains the likely values of a parameter.

Point Estimation

One does not always have to estimate the complete distribution of a random variable X

Often, interest is in specific characteristics of the distribution, such as the population average μ

Definition: A point estimate is a single numerical value used to estimate the corresponding population parameter.

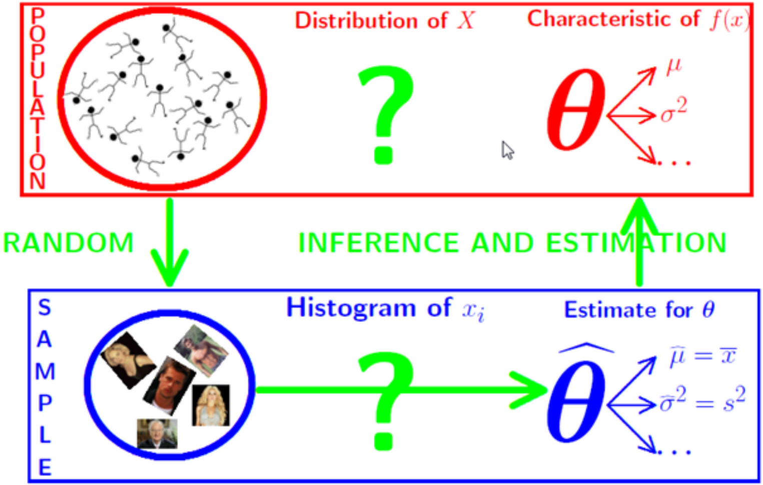
A parameter is a numerical descriptive measure of a population whereas statistic is a numerical descriptive measure of a sample ($\overline{X}$ is an example of a statistic).

To each population parameter, there are corresponding sample statistic.

Estimate Population parameter		with sample statistics
Mean	μ	$\hat{\mu} = \overline{x}$
Variance	σ^2	$\hat{\sigma}^2 = s^2$
Proportion	P	$\hat{P} = p$
Difference	$\mu_1 - \mu_2$	$\overline{x}_1 - \overline{x}_2$

Note that it is very unlikely that the estimate is identical to the parameter it is estimating

How close the estimate will be to the true value depends on various aspects.


POPULATION
Distribution of X
Characteristic of $f(x)$
?
θ
μ
σ^2
$\dots$
RANDOM
INFERENCE AND ESTIMATION
SAMPLE
Histogram of x_i
Estimate for θ
?
$\widehat{\theta}$
$\widehat{\mu} = \overline{x}$
$\widehat{\sigma}^2 = s^2$
$\dots$

- Suppose interest is in the estimation of some characteristic Θ of the distribution of a specific random variable X
- Based on a random sample, an estimate for Θ can be obtained, for example:

	Population parameter (never observed)	Estimate from sample (observed)
mean	$\theta = \mu$	$\hat{\mu} = \overline{x}$
variance	$\theta = \sigma^2$	$\hat{\theta} = s^2$
In general	θ	$\hat{\theta}$

- The estimate $\hat{\theta}$ is calculated from the observed data, hence the resulting value for $\hat{\theta}$ completely depends on the sample that was drawn from the population

Definition of terms

Confidence Interval: An interval estimate with a specific level of confidence.

Confidence Level: The percent of the time that the true value will lie in the interval estimate given.

Consistent Estimator: An estimator which gets closer to the value of the parameter as the sample size increases.

Degrees of Freedom: The number of data values which are allowed to vary once a statistic has been determined.

Interval Estimate: A range of values used to estimate a parameter.

Point Estimate: A single value used to estimate a parameter.

Relatively Efficient Estimator: The estimator for a parameter with the smallest variance.

Unbiased Estimator: An estimator whose expected value is the value of the parameter being estimated

Point and Interval estimation of the population mean: μ

Point Estimation about population mean μ

Another term for statistic is point estimate.

For instance, sum of X_i over n is the point estimator used to compute the estimate of the population mean μ.

That is $\overline{X} = \frac{\sum_i X_i}{n}$ is a point estimator of the population mean.

Confidence interval estimation of the population mean

The confidence level is the probability that the value of the parameter falls within the range specified by the confidence interval surrounding the statistic.

There are different cases to be considered to construct confidence intervals.

Case I: If sample size is large or if the population is normal with known variance.

Recall the Central Limit Theorem, which applies to the sampling distribution of the mean of a sample.

Consider samples of size n drawn from a population, whose mean is μ and standard deviation is σ with replacement and order is important. The population can have any frequency distribution. The sampling distribution of $\overline{X}$ will have a mean $\mu_{\overline{X}} = \mu$ and a standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$ and approaches a normal distribution as n gets large.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\begin{aligned} \Rightarrow \mu &= \overline{X} \pm Z\sigma/\sqrt{n} \\ &= \overline{X} \pm \epsilon \quad \text{where} \quad \epsilon \quad \text{is a measure of error} \end{aligned}$$

$\Rightarrow \epsilon = Z\sigma/\sqrt{n}$

Parameter = Statistic ± Its Error

For the interval estimator to be good the error should be small. How it be small?

By making n large

Small variability

Taking Z small

To obtain the value of Z, we have to attach this to a theory of chance. That is, there is an area of size $1-\alpha$

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1-\alpha$$

where α The parobability that the parameter lies outside the interval

$Z_{\alpha/2}$ the standard normal variable to the right of which $\alpha/2$ probability lies i.e $P(Z > Z_{\alpha/2}) = \alpha/2$

$$\Rightarrow P\left(-Z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(\overline{X} - Z_{\alpha/2}\sigma/\sqrt{n} < \mu < \overline{X} + Z_{\alpha/2}\sigma/\sqrt{n}\right) = 1-\alpha$$

$$\left(\overline{X} - Z_{\alpha/2}\sigma/\sqrt{n}, \overline{X} + Z_{\alpha/2}\sigma/\sqrt{n}\right) \text{ is A } 100(1-\alpha)\%\text{CI for } \mu$$

Case II: If the sample size is large and the variance σ^2 is unknown

Usually σ^2 is not known, in that case we estimate by its point estimator

$$\left(\overline{X} - Z_{\alpha/2} S/\sqrt{n}, \overline{X} + Z_{\alpha/2} S/\sqrt{n}\right) \text{ is A } 100(1-\alpha)\% \text{ CI for } \mu$$

Case III: If the sample size is small and the population variance σ^2 is unknown

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} \quad \text{has } t \text{ distribution on with } n-1 \text{ degree of freedom}$$

$$\left(\overline{X} - t_{\alpha/2} S/\sqrt{n}, \overline{X} + t_{\alpha/2} S/\sqrt{n}\right) \text{ is A } 100(1-\alpha)\% \text{ CI for } \mu$$

Examples

From a normal sample of size 25 a mean of 32 was found .Given that the population standard deviation is 4.2. Find A 95% confidence interval for the population mean.

A drug company is testing a new drug which is supposed to reduce blood pressure. From the six people who are used as subjects, it is found that the average drop in blood pressure is 2.28 points, with a standard deviation of 0.95 points. What is the 95% confidence interval for the mean change in pressure?

Answer

1. Use Case I

$\overline{X} = 32, \sigma = 4.2 \;\; 1 - \alpha = 0.95 \Rightarrow \alpha = 0.05, \alpha = 0.05/2 = 0.025$
$\Rightarrow Z_{\alpha/2} = 1.96$ from table
The required interval will be $\overline{X} \pm Z_{\alpha/2}\sigma/\sqrt{n} = 32 \pm 1.96 * 4.2/\sqrt{25}$

$$32 \pm 1.65 = (30.35, 33.65)$$

2. Use Case III

$\overline{X} = 2.28, S = 0.95 \;\; 1 - \alpha = 0.95 \Rightarrow \alpha = 0.05, \alpha = 0.05/2 = 0.025$
$\Rightarrow t_{\alpha/2} = 2.571$ with df=5 from table

The required interval will be $\overline{X} \pm t_{\alpha/2}\sigma/\sqrt{n}$

$$2.28 \pm 2.571 * 0.96/\sqrt{6}$$

$$2.28 \pm 1.008 = (1.28, 3.28)$$

Hypothesis Testing

Hypothesis; is statement about a population developed for the purpose of testing.

Hypothesis testing: is a procedure based on sample evidence and probability to determine theory whether the hypothesis is a reasonable statement.

Test statistic: is a value determined from a sample information, used to determine whether to reject the null hypothesis or not.

Decision rule: is a statement of the condition under which the null hypothesis is rejected and the conditions under which it is not rejected.

Critical region: A region where the null hypothesis is rejected

There are research hypotheses and statistical hypotheses.

Research hypothesis: is the supposition or conjecture that motivates the research.

There are two statistical hypotheses involved in hypothesis testing.

Null hypothesis:

- It is the hypothesis to be tested.
- It is the hypothesis of equality or the hypothesis of no difference.
- It is denoted by H_0.

Alternative hypothesis:

- It is the hypothesis available when the null hypothesis has to be rejected.
- It is the hypothesis of difference.
- It is denoted by H_1 or H_a.

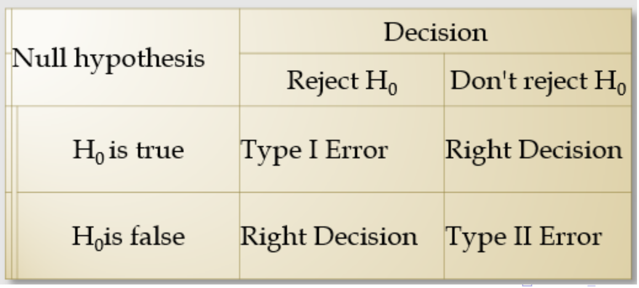
Level of significance: The level of significance α, is the probability, in reality, the probability of rejecting the null hypothesis

Significance and errors : When the computed value of the test statistic falls in the rejection region it is said to be significant
We select a small value of α such as 0.1, 0.05 or 0.01 to make the probability of rejecting a true null hypothesis small.

Types of errors:

Testing hypothesis is based on sample data which may involve sampling errors.

The following table gives a summary of possible results of any hypothesis test:



Null hypothesis	Decision	
	Reject H_0	Don't reject H_0
H_0 is true	Type I Error	Right Decision
$H_0$is false	Right Decision	Type II Error

Type I error: Rejecting the null hypothesis when it is true. It is denoted by α

Type II error: Accepting the null hypothesis when it is false. It is denoted by β

A p-value is the probability that the computed value of a test statistic is at least as extreme as a specified value of the test statistic when the null hypothesis is true. Thus, the p-value is the smallest value of for which we can reject a null hypothesis.

Procedure for hypothesis testing

1. Data
2. Assumptions
3. Hypotheses
4. Test statistic
   a. Distribution of test statistic
   b. Decision rule
5. Calculation of test statistic
6. Statistical decision
7. Conclusion

Hypothesis Testing of a Single Population Mean

- A hypothesis about a population mean can be tested when sampling is from any of the following.
  - A normally distributed population–variances known
  - A population that is not normally distributed and variance unknown (assuming $n > 30$; the central limit theorem applies)
  - A normally distributed population–variances unknown and small sample size (using t-distribution)
- Suppose the assumed or hypothesized value of μ is denoted by μ_0 , then one can formulate two sided (1) and one sided (2 and 3) hypothesis as follows:

1. $H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$
2. $H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$
3. $H_0 : \mu \geq \mu_0 \quad \text{vs} \quad H_1 : \mu < \mu_0$

CASES

Case 1: When sampling is from a normal distribution with σ^2 known: The relevant test statistic is

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

After specifying α we have the following regions (critical and acceptance) on the standard normal distribution corresponding to the above three hypothesis

H_0	Reject H_0 if	Accept H_0 if
$\mu = \mu_0$	$\lvert Z_{call}\rvert > Z_{\alpha/2}$	$\lvert Z_{call}\rvert < Z_{\alpha/2}$
$\mu \leq \mu_0$	$Z_{call} < -Z_{\alpha}$	$Z_{call} > -Z_{\alpha}$
$\mu \geq \mu_0$	$Z_{call} > Z_{\alpha}$	$Z_{call} < Z_{\alpha}$

Case 2: When sampling is from a normal distribution with σ^2 unknown and small sample size: The relevant test statistic is

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

After specifying α we have the following regions (critical and acceptance) on the t-distribution

H_0	Reject H_0 if	Accept H_0 if
$\mu = \mu_0$	$\lvert t_{call} \rvert > t_{\alpha/2}$	$\lvert t_{call} \rvert < t_{\alpha/2}$
$\mu \leq \mu_0$	$t_{call} < -t_{\alpha}$	$t_{call} > -t_{\alpha}$
$\mu \geq \mu_0$	$t_{call} > t_{\alpha}$	$t_{call} < t_{\alpha}$

Case III: When sampling is from a non- normally distributed population or a population whose functional form is unknown

If a sample size is large one can perform a test hypothesis about the mean by using:

Test statistic is $Z_{cal} = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$, if σ^2 is known

$$= \frac{\overline{X}-\mu_0}{S/\sqrt{n}}, \text{if } \sigma^2 \text{ is unknown}$$

The decision rule is similar to case I

When a true null hypothesis is rejected, it causes a Type I error whose probability is α.

When a false null hypothesis is not rejected, it causes a Type II error whose probability is designated by β.

A Type I error is considered to be more serious than a Type II error.

	Condition of null hypothesis	
Possible Action	True	False
Fail to reject H_0	Correct $(1-\alpha)$	Type II error (β)
Reject H_0	Type I error α	Correct $(1-\beta)$

p -value is a probability that the result is as extreme or more extreme than the observed value if the null hypothesis is true.

If the p value is less than or equal to α, we reject the null hypothesis, otherwise we do not reject the null hypothesis.

Example: Sampling from a normally distributed population with variance known. A simple random sample of 10 people from a certain population has a mean age of 27. Can we conclude that the mean age of the population is not 30? The variance is known to be 20. Let $\alpha = 0.05$

Solution:

a. Data: $n = 10, \quad \sigma^2 = 20, \quad \overline{X} = 27, \quad \alpha = 0.05$
b. Assumption: population are normally distributed
c. Hypothesis $H_0 : \mu = 30, \quad H_A : \mu \neq 30$
d. Test statistic

$$\begin{aligned} Z &= \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{27 - 30}{\sqrt{20/10}} \\ &= \frac{-3}{1.4142} = -2.12 \end{aligned}$$

e. Decision: Because of the structure of H_0 it is a two tail test. Therefore, reject H_0 if $z \leq -1.96$ or $z \geq 1.96$

f. Discussion: we reject the null hypothesis because $z = -2.12$ which is in the rejection region. the value is significant at $\alpha = 0.05$ level.

g. Conclusions: we conclude that at 95% level of significance we have enough evidence to conclude that the mean age of the population is not 30. $p.value = 0.034$

Confidence interval

$$\overline{X} \pm Z\sigma\sqrt{n}$$

$$27 \pm 1.96\sqrt{20/10}$$

$$27 \pm 1.96(1.4142)$$

$$(24.228, 29.772)$$